Phd Thesis

University of Copenhagen

Faculty of Science

Under the double degree agreement with

Università di Milano-Bicocca

Chemometrics approaches for the automatic analysis of metabolomics GC-MS data

Author:

Giacomo Baccolo

Supervisors:

Prof. Davide Ballabio

Prof. Rasmus Bro

SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of
Earth and Environmental Sciences

PhD program in Chemical, Geological and Environmental Sciences, Cycle XXXIV
Curriculum in Chemical Sciences

Under the double degree agreement with the
University of Copenhagen

Chemometrics approaches for the automatic analysis of metabolomics GC-MS data

Baccolo Giacomo
753583

Tutor: Professor Marco Emilio Orlandi

Supervisor: Professor Davide Ballabio
Supervisor: Professor Rasmus Bro

Coordinator: Professor Marco Giovanni Malusà

ACADEMIC YEAR 2020-2021

UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD thesis 2022 - Giacomo Baccolo

Chemometrics approaches for the automatic analysis of metabolomics GC-MS data

Acknowledgements

This thesis has been submitted to the PhD School of the Faculty of Science, University of Copenhagen as well as to the PhD School of Chemical, Geological and Environmental Sciences, University of Milano-Bicocca, in fulfilment of the requirements to obtain the PhD degree

Contents

List of Figures

List of Tables

Publications and deliverable

Publications

Published

- A MATLAB toolbox for multivariate regression coupled with variable selection.
  Consonni, V., **Baccolo, G.**, Gosetti, F., Todeschini, R., Ballabio, D. (2021).
  Chemometrics and Intelligent Laboratory Systems, 213, 104313

- From untargeted chemical profiling to peak tables–A fully automated AI driven approach to untargeted GC-MS.
  Baccolo, G., Quintanilla-Casas, B., Vichi, S., Augustijn, D., Bro, R. (2021)
  TrAC Trends in Analytical Chemistry, 145, 116451

Near to submission

- Machine learning approaches for the classification of resolved elution profiles.
  Baccolo, G., Yu, H., Valsecchi, C., Ballabio, D., Bro, R.

- PARASIAS: a new method for higher-order tensor analysis with shifting profiles.
  Yu, H., **Baccolo, G.**, , Bro, R., Neal B. Gallagher, N.B.

Deliverables

- Regression toolbox (for MATLAB)
  avaible at: https://michem.unimib.it/download/matlab-toolboxes/

- AutoDise GUI
  available on request, writing to: g.baccolo1@campus.unimib.it

List of Abbreviations

Adam	**Ada**ptative **M**ovement estimator
ALS	**A**lternative **L**east **S**quare
ANN	**A**rtificial **N**eural **N**etwork
AUC	**A**rea **U**nder (the) **C**urve
BILSTM	**BI**linear **L**ong **S**hort **T**ime **M**emory
CNN	**C**onvolutional **N**eural **N**etwork
CORCONDIA	**COR**e **CON**tor DIAgonstic
COW	**C**orrelated **O**ptimized Warping
Da	**Da**lton
EI	**E**lectron **I**onization
EIC	**E**xtracted **I**on **C**romatogram
FPR	**F**alse **P**ositive **R**ate
GC-MS	**G**as **C**hromatography - **M**ass **S**pectrometry
GUI	**G**raphical **U**ser **I**nterface
KNN	**K** **N**earest **N**neighbors
LC-MS	**L**iquid **C**hromatography - **M**ass **S**pectrometry
LSTM	**L**ong **S**hort **T**ime **M**emory
MCR	**M**ultivariate **C**urve **R**esolution
NER	**N**on **E**rror **R**ate
NIPALS	**N**on **I**terative **PA**rtial **L**east **S**quare
NMR	**N**uclear **M**agnetic **R**esonance
PARADISe	**PAR**AFAC2 based **D**econvolution and Identification System
PARAFAC	**PAR**allel **FAC**tor analysis

PARAFAC2	**PAR**allel **FAC**tor analysis 2
PCA	**P**rincipal **C**omponent **A**nalysis
PPP	**P**arallel **P**roportional **P**rofiles
RMSProp	**R**oot **M**ean **S**quared **Prop**agation
RNN	**R**ecurrent **N**eural **N**etwork
ROC	**R**eciving **O**perating **C**haracteristic
SGD	**S**tocastic **G**radient **D**escend
SVD	**S**ingular **V**alue **D**ecomposition
TPR	**T**rue **P**ositive **R**ate

Sommario

Giacomo Baccolo

Chemometrics approaches for the automatic analysis of metabolomics GC-MS data

La metabolomica, che consiste nella identificazione di tutti i metaboliti presenti all'interno dei campioni biologici analizzati, è un approccio affascinante e ampiamente applicato in diversi campi di ricerca quali: identificazione di biomarcatori, sviluppo di nuovi farmaci scienze alimentari e ambientali.
La metabolomica è strettamente legata alla capacità di tecniche analitiche fra queste una delle più applicate è la gas cromatografia accoppiata alla spettrometria di massa. Moderne piattaforme analitiche possono generare centinaia di migliaia di spettri, rilevando una quantità impressionante di molecole distinte. Nonostante i progressi tecnici raggiunti sul lato sperimentale, o a causa di questi, la conversione dei segnali misurati dagli strumenti in informazioni utili non è un passaggio scontato in studi metabolomici. Per ogni composto identificato, l'obbiettivo è ottenere la concentrazione relativa tra tutti i campioni analizzati e lo spettro di massa associato al composto utilizzato poi per l'identificazione della molecola stessa. I segnali ottenuti da strumenti GC-MS sono complessi e i software disponibili per l'analisi dei dati sperimentali sono stati ripetutamente indicati come una fonte importante di incertezza, limitando fortemente sia la quantità che la qualità delle informazioni estratte. La maggior parte degli approcci, o almeno i più applicati, si basano sull'analisi univariata dei dati, considerando ogni campione separatamente dagli altri e richiedono l'impostazione di diversi parametri da parte dell'operatore, influenzando il risultato dell'analisi.
In questa tesi è descritto un nuovo approccio, chiamato AutoDise, per l'analisi dei dati

GC-MS. L'elaborazione dei segnali sperimentali si basa su PARAFAC2. PARAFAC2 è un modello che scompone dati multidimensionali, discriminando tra i diversi segnali nei campioni. L'efficacia di PARAFAC2 nell'estrarre segnali chimici dai dati GC-MS è stata ampiamente dimostrata. Grazie alle sue proprietà intrinseche, PARAFAC2 non ha quasi bisogno che i dati siano pretrattati e non richiede di impostare parametri, mentre software utilizzati in questo ambito richiedono di definire diversi parametri e un laborioso pretrattamento dei dati, è quindi necessario l'intervento di un utente esperto, inoltre la riproducibilità dei risultati è limitata, dipendendo da come i parametri sono stati scelti per l' analisi.

Tuttavia, il fitting di modelli PARAFAC2 coinvolge diverse fasi ed è necessario un esperto analista per l'analisi e l'interpretazione dei modelli. AutoDise è un sistema esperto in grado di gestire tutti i passaggi riguardanti la modellazione e di generare una tabella dei picchi in cui ogni composto è identificato in modo univoco, con risultati completamente riproducibili. Questo è possibile grazie alla combinazione di diverse strumenti diagnostici e grazie all' applicazione di modelli d'intelligenza artificiale. Le prestazioni dell'approccio sono state testate su un complesso dataset di oli d'oliva ottenuto tramite analisi GC-MS. I dati sono stati analizzati sia manualmente, da utenti esperti, sia automaticamente con il metodo AutoDise proposto e le tabelle dei picchi risultanti sono state confrontate. I risultati mostrano che AutoDise supera l'analisi manuale sia in termini di numero di composti identificati che di qualità dell'identificazione e della quantificazione. Inoltre, è stata sviluppata una GUI dedicata per rendere l'algoritmo più accessibile a persone non esperte nel linguaggio di programmazione. Nella tesi è incluso un tutorial che mostra le caratteristiche principali e come utilizzare l'interfaccia grafica.

Un'altra parte importante del progetto è stata dedicata al test e allo sviluppo di nuove reti neurali artificiali da implementare nel software AutoDise per rilevare quali componenti PARAFAC2 stanno fornendo informazioni chimicamente utili. A tal fine, più di 170.000 profili sono stati etichettati manualmente, al fine di addestrare, validare e testare una rete neurale convoluzionale e una rete bilineare con memoria a breve termine e un modello k-nearest neighbour. I risultati suggeriscono che le reti di deep learning possono essere efficacemente applicate per la classificazione automatica dei profili cromatografici.

Resumé

Giacomo Baccolo

Chemometrics approaches for the automatic analysis of metabolomics GC-MS data

Metabolomis, der sigter mod en omfattende karakterisering af metabolitterne i en prøve, er en tiltalende og bredt anvendt videnskabelig metode, der anvendes inden for adskillige forskningsområder, såsom identifikation af biomarkører, lægemiddelopdagelse, fødevarevidenskab og miljøvidenskab. Ofte er de analyserede prøver karakteriseret ved komplekse matricer, f.eks. biologisk væv, fødevarer eller jordprøver. Metabolomics er tæt forbundet med analytiske teknikker, og en af de mest anvendte er gaskromatografi i forbindelse med massespektrometri. Moderne analytiske platforme kan generere hundredtusindvis af spektre og påvise en imponerende mængde af forskellige molekyler. På trods af, eller på grund af, de tekniske fremskridt, der er opnået på den eksperimentelle side, er oversættelsen af de signaler, der måles af instrumenterne, til letanvendelige oplysninger, stadig en flaskehals inden for metabolomics. For hver identificeret forbindelse ønskes den relative koncentration på tværs af de analyserede prøver samt det tilhørende massespektrum. Signalerne fra GC-MS-instrumenter er komplekse, og den software, der er til rådighed til analyse af de eksperimentelle data, er gentagne gange blevet angivet som en væsentlig kilde til usikkerhed, hvilket i høj grad begrænser både kvantiteten og kvaliteten af de ekstraherede oplysninger. De fleste metoder, eller i hvert fald de mest anvendte, er baseret på univariat analyse af data, hvor hver prøve betragtes separat fra de andre. Ofte kræves en besværlig manuel indsats for at indstille de ofte mangeparametre, der påvirker analyseresultatet. Denne afhandling omhandler præsentationen af en ny metode kaldet AutoDise til analyse af GC-MS-data. Det mest kritiske trin består

i at udtrække bidrag fra de enkelte kemiske komponenter fra de eksperimentelle data ved hjælp af modellen PARAFAC2. PARAFAC2-modelleringsmetoden dekomponerer de multilineære data og skelner mellem de forskellige signaler på tværs af prøverne. PARAFAC2's effektivitet med hensyn til at udtrække meningsfulde kemiske signaler fra GC-MS-data er blevet demonstreret i vid udstrækning. På grund af sine egenskaber har PARAFAC2 næsten intet behov for forbehandling af data, og der kræves ingen kritiske indstillinger, hvorimod andre metoder kræver indstilling af flere parametre og en besværlig forbehandling af dataene, hvilket kræver at brugeren er yderst erfaren. . Tilpasning af PARAFAC2-modeller involverer imidlertid forskellige faser, og der kræves en dygtig tensoranalytiker til analyse og fortolkning af modellerne. AutoDise er et ekspertsystem baseret på statistisk diagnostik og kunstig intelligens, som kan tage sig af alle modelleringsaspekter og generere en peaktabel, hvor hver kemisk forbindelse er identificeret og kvantificeret, og hvor resultaterne er reproducerbare. Metodens ydeevne er blevet testet på et komplekst datasæt af jomfruolivenolier af forskellige kvalitetsklasser, hvis aroma profiler blev målt ved hjælp af fastfasemikroekstraktion - GC/MS. Dataene er blevet analyseret både manuelt af erfarne brugere og automatisk med den foreslåede AutoDise-metode, og de resulterende peaktabeller er blevet sammenlignet. Resultaterne viste, at AutoDise overgår den manuelle analyse både med hensyn til antallet af identificerede forbindelser og kvaliteten af identifikationen og kvantificeringen. Desuden er der udviklet en dedikeret GUI for at gøre algoritmen mere tilgængelig for personer, der ikke har kendskab til programmeringssprog. Afhandlingen indeholder en vejledning, der viser de vigtigste funktioner og viser, hvordan man bruger den grafiske grænseflade. En anden vigtig del af afhandlingen var afsat til test og udvikling af nye kunstige neurale netværk, der skal implementeres i AutoDise-softwaren til at registrere hvilke PARAFAC2-komponenter, der giver kemisk nyttig information. Med henblik herpå er mere end 170 000 profiler blevet mærket manuelt med henblik på at træne, validere og teste et konvolutionelt neuralt netværk og et bilineært langtids- og korttidshukommelsesnetværk samt en k-nærmeste nabomodel. Resultaterne tyder på, at deep learning-netværk effektivt kan anvendes til automatisk klassificering af kromatografiske profiler.

Abstract

Giacomo Baccolo

Chemometrics approaches for the automatic analysis of metabolomics GC-MS data

Metabolomics, aiming at the comprehensive characterization of the metabolites in a sample, is an appealing and widely applied scientific approach applied in several research fields, such as biomarker identification, drug discovery, food science and environmental science. Often, the analyzed samples are characterized by complex matrices, such as biological tissues, foods or soil samples. Metabolomics is closely related to analytical techniques and one of the most applied is Gas Chromatography connected to Mass Spectrometry. Modern analytical platforms can generate hundreds of thousands of spectra, detecting an impressive amount of distinct molecules. Despite, or because, the technical progress achieved on the experimental side, the translation of the signals measured by the instruments into easily expendable information is still a major bottleneck in metabolomics. For each identified compound, it is desired to have the relative concentration across the analyzed samples as well as the associated mass spectrum. The signals from GC-MS instruments are complex and the software available for the analysis of the experimental data have been repeatedly indicated as a major source of uncertainty, strongly limiting both the quantity and the quality of the extracted information. Most of the approaches, or at least the most applied, are based on the univariate analysis of the data, considering each sample separately from the others and requiring laborious manual efforts for the setting of several parameters that affect the result of the analysis. This thesis deals with the presentation of a new approach called AutoDise for the analysis of GC-MS data. The most critical step, that consist in the extraction of the pure contributions from the experimental

data, is performed by means of PARAFAC2. The PARAFAC2 modelling approach decomposes the multilinear data, discriminating among the different signals across the samples. The efficacy of PARAFAC2 to extract meaningful chemical signals from GC-MS data has been widely demonstrated. Because of its intrinsic properties, PARAFAC2 has almost no need for data preprocessing, no critical settings are required, whereas other approaches need to set several parameters and a laborious pretreatment of the data, requiring an extremely skilled user and dramatically reducing results reproducibility. However, fitting PARAFAC2 models involves different phases and a skilled tensor analyst is required for the analysis and interpretation of the models. AutoDise is an expert system based on statistical diagnostics and Artificial Intelligence, which is able to take care of all the modeling aspects and to generate a peak table where each compound is univocally identified and fully reproducible results. The performance of the approach has been tested on a complex dataset of virgin olive oils with different quality grades, whose volatile profiles were obtained by solid-phase microextraction - GC/MS. The data have been analyzed both with a commercial software, by experienced users, and automatically with the proposed AutoDise method and the resulting peak tables have been compared. The results showed that AutoDise overcome the manual analysis both in terms of number of identified compounds and in quality of the identification and quantification.

Moreover, a dedicated GUI has been developed to make the algorithm more accessible to people not skilled in programming language. A tutorial is included in the thesis, showing the main features and how to use the graphical interface.

Another important part of this thesis is devoted to the test and development of new artificial neural networks to be implemented in the AutoDise software for detecting which PARAFAC2 components that are providing chemically useful information. To this end, more than 170,000 profiles have been manually labeled, in order to train, validate and test a convolutional neural network and a bilinear long short term memory network and a k-nearest neighbour model. The results suggest that deep learning networks can effectively be applied for the automatic classification of the chromatographic profiles.

Chapter 1

Introduction and aim of the thesis

In 1951 Roger Williams, biochemist at the University of Texas, introduced the concepts that nowadays are considered as the basis of omics and related fields such as systems biology and personalized medicine, where omics tools are massively applied (Rattray et al., 2014; Huan et al., 2017; Rosato et al., 2018; Jacob et al., 2019).

> Generalizations with respect to human beings are very much to be desired, but unless the generalizations are backed by adequate information, they are premature and superficial and liable to be incapable of substantiation. The writer envisaged and still envisages the time when real people will be studied from any and every angle and when no promising discipline will be ignored in connection with such studies (Williams and Berry, 1951).

In particular, Williams and his colleagues were focused on *metabolic patterns*. Analyzing body fluids, they demonstrated that specific conditions, e.g., alcoholism or schizophrenia, were characterized by specific metabolites. He defined *metabolic patterns* as follow:

> Insofar as the total picture of the metabolism of an individual (including the chemical processes in each and every organ and tissue and their effects on each other) is distinctive, it constitutes his metabolic pattern (Williams and Berry, 1951).

In more recent years, this definition would perfectly fit metabolomics. Since Williams' pioneering work, the development of metabolomics or metabolic profiling has

been intrinsically linked to a comprehensive overview of the analyzed samples, aiming to identify as many metabolites as possible, ideally all of them.

Formally a metabolite is an intermediate or an end product in biosynthetic or degradative pathways of a given organism (Cox and Nelson, 2008), as such metabolomics is focused on relatively small molecules and usually the molecular weight does not exceed 1000-1500 Da.

Common metabolomics applications range from bio markers identification (Johnson, Ivanisevic, and Siuzdak, 2016) to drug discovery (Wishart, 2016), however the applications where this philosophy has been embraced are countless, ranging from food fraud detection (Cubero-Leon, Peñalver, and Maquet, 2014), to environmental risk assessment (Bundy, Davey, and Viant, 2009). Moreover, metabolic profiling is increasingly applied in regulatory contexts such as industrial chemistry and food safety (European Chemicals Agency, 2016; European Food Safety, 2014).

The number of molecules that can be identified with a metabolomics experiment obviously depends on the samples analyzed, but it is impressive. For example, the metabolites in an organism are estimated between 1000 and 40000, depending on the species (Alseekh and Fernie, 2018). Nucleotides, vitamins, organic acids, lipids, antibiotics are just a small fraction of the molecules that make up the metabolome. Metabolites are characterized by a rapid turnover and the broad dynamic range of abundance of the different molecules (Alseekh et al., 2021). Modern analytical platforms can detect hundreds or even thousands of different molecules, apparently making the vision of metabolomics studies a reality (May and McLean, 2016; Tauler and Parastar, 2018). However, a metabolomic experiment is a multistep process (Goodacre et al., 2007), where the analysis of the samples is just one of the first points.

A common pipeline for a metabolomic experiment consists of:

1. Experimental design;
2. Sample collection;

3. Data acquisition;
4. Data preprocessing;
5. Metabolite/molecules identification and quantification;
6. Statistical data analysis;
7. Interpretation.

Experimental design includes selecting which analytical platform to use, the type and the number of samples, the number of experiments, in relation to the specific aim, e.g., to test a biological hypothesis or to discover new biomarkers.
After that, during the samples collection they are prepared for analysis through the analytical platform chosen. This usually entails an extraction or a derivatization step.
Next, the data generated by the analytical platform has to be preprocessed in order to extract the relevant information from the data, i.e., identify the signals related to the chemical compounds.
The extracted signals are used to quantify and identify chemical compounds. This step leads to the peak table where for each sample the amount of each compound is reported and its structure is identified.
The statistical data analysis of the peak table and interpretation steps aim at answering the scientific question at the base of the experiment through (univariate/multivariate) statistical tools. As such the last two steps are the core of all the metabolic profiling experiments and intrinsically depend on the results of the previous phases.

While on the experimental side major progress has been made, data preprocessing and molecule identification have been repeatedly reported as the main bottleneck in metabolomics experiments (Dunn et al., 2013; Sévin et al., 2015; Schrimpe-Rutledge et al., 2016; Chaleckis et al., 2019), despite the number of tools developed by the scientific community to handle this step (Stein, 1999; Tautenhahn et al., 2012; Tsugawa et al., 2015; Domingo-Almenara et al., 2016). This is reflected in issues such as:

- The time consumption for the analysis of the instrumental signals (Koek et al., 2011);

- The not effective translation of the experimental data into new knowledge (Clarke et al., 2008);
- The low reliability of the analysis performed on the peak table, due to the false discovery rate of the compounds; (Schrimpe-Rutledge et al., 2016);
- The non-robustness, e.g., in the sense that different users often do not obtain similar results or different software packages do not achieve the same results

The main goal of this project was to develop a truly automated algorithm, called AutoDise, for the automate analysis of gas chromatography - mass spectrometry data, one of the most applied techniques for metabolomics experiments, obtained from metabolomics experiments. The algorithm is based on the combination of chemometric, statistical and deep learning approaches to automatically resolve, extract, quantify and identify the signals of the chemical compounds measured by the instrument. The output consists of a non redundant peak table, thus AutoDise automatically handles the data preprocessing and metabolite/molecule identification and quantification steps with no actions required by the user. The main advantages we hope and hypothesize to gain compared to the current state-of-the-art are:

- The absence of interaction with user, reducing the variability of the resulting peak table;
- Allowing to reliably detect many more compounds, e.g., compounds that are not visually apparent or are suffering from severe coelution.

The rest of this thesis is organized as follow:

- The analytical approaches applied in metabolic profiling, particularly focused on Gas Chromatography - Mass Spectrometry, are described in Chapter 2;
- An overview of the available approaches for the analysis of GC-MS data is given in Chapter 3;
- The methods employed during the project are described in Chapter 4;
- In Chapter 5 the results obtained concerning AutoDise are shown. The overall method, a case study to test the algorithm performances, the GUI implemented

to support the use of the algorithm and the results for the application of ANN for the classification of resolved elution profiles are described.

- The conclusions and the future perspectives are discussed in Chapter 6

Chapter 2

Analytical platforms in metabolic profiling

Regardless of the specific field of application, metabolomics is strictly related to the development of analytical techniques for the detection of molecules. Starting from the paper chromatography applied by Williams and his colleagues, where the experimental results were mainly qualitative (Williams and Berry, 1951), nowadays modern analytical platforms can analyze impressive number of samples and discriminate hundreds of molecules.

In the context of metabolomics and related fields, there are two main analytical approaches:

- NMR;
- hyphenated chromatography.

Both these approaches have strengths and weaknesses.

NMR is characterized by high reproducibility, low cost and time for the analysis of the sample. Instead, hyphenated chromatography is by far more sensitive, detecting more metabolites even at very low concentration (Emwas, 2015); Wishart, 2008 Markley et al., 2017 and Wishart, 2019 are suggested for an excellent overview about NMR in the context of metabolomics.

This thesis is focused on the analysis of hyphenated chromatography data. Hyphenated chromatography generally refers to the online combination of two separate analytical

techniques connected by appropriate interfaces (Hirschfeld, 1980). In the context of metabolomics, most often, the chromatographic system is connected to a mass spectrometer. The connection of multiple chromatographic systems and mass spectrometers is also possible, leading to the so called hypernation (Wilson and Brinkman, 2003).

Samples as food, oils, body fluids, cells are characterized by complex matrices and include several different chemical compounds with significant differences in terms of abundance from molecule to molecule. The coupling of a chromatographic method and a mass spectrometer takes advantage of both the methods. Chromatography separates pure or nearly pure bands of chemical components in a mixture. Mass spectrometry gives selective information for identification using standards or library spectra.

2.1 Targeted and untargeted approaches

In metabolomics the two main experimental approaches are the targeted approach and the untargeted approach. The targeted approach refers to the analysis of a subset of defined metabolites/molecules. Usually, the subset is selected according to the experimental hypothesis. Defining *a priori* the subset of molecules allows an appropriate setting of the experimental procedures, but the analysis are typically limited to a small number of molecules. However, lead by technological advances, the subset of metabolites that can be analyzed simultaneously has steadily increased over the recent years (Roberts et al., 2012), up to a hundred.

The untargeted approach refers to the measurement of all the molecules within the analyzed sample, rather than just a small, predefined subset (Schrimpe-Rutledge et al., 2016). The analysis of the results and the comparison of samples can lead to new hypotheses that require *ad hoc* experiment to be tested.

The two approaches can be seen as complementary to each other: the hypothesis tested with a targeted analysis comes from observation based on the experiments performed with an untargeted approach (Diederen et al., 2021). The focus in this thesis is on the analysis of data obtained with an untargeted approach.

2.2 Introduction to chromatography

The term chromatography derives from the Greek words $\chi\rho\omega\mu\alpha$ (*chroma*) and $\gamma\rho\alpha\varphi\omega$ (*grapho*), it was introduced by the Russian botanist Mikhail Tswett, who invented this method. He packed a column of glass with calcium carbonate and inserted a sample of plant extract at the top of the column. Then he filled the column with a mixture of petrol ether and ethanol. The flow of the mixture caused the plant pigments, contained in the extracts, to separate along the column with three different colored bands, hence the name chromatography.

Since then, a number of different chromatographic system have been developed, nonetheless all of them have in common the flow of a mixture of analytes (the plant extract in Tswetts' experiment) through a stationary phase (calcium carbonate) carried by a mobile phase (the mixture of petrol ether and ethanol).
The first classification for chromatographic systems is column chromatography, where the stationary phase is placed in a column, or planar chromatography where the stationary phase is hold on the surface of a flat plate or in a paper. In metabolomics and metabolic profiling column chromatography is the gold standard.

Column chromatography can be classified depending on the state of the mobile phase: liquid and gas chromatography. This thesis is mainly focused on gas chromatography but some of the considerations in this paragraph apply in general for column chromatography.
The column is a thin tube with an adsorptive material (the stationary phase), coated on either inert solid particles (usually LC) or the walls of the tube (usually GC); the mobile phase, gaseous or liquid, is introduced in the column from the head. The sample, made of a mixture of different compounds, is injected in the head of the column. Fresh mobile phase is added, forcing the movement trough the column of the compounds. This process is called **elution** and the fresh mobile phase is the **eluent**.

The fundamental mechanism that separates the molecules of a sample is based on the different distribution constants between the mobile and stationary phase of the

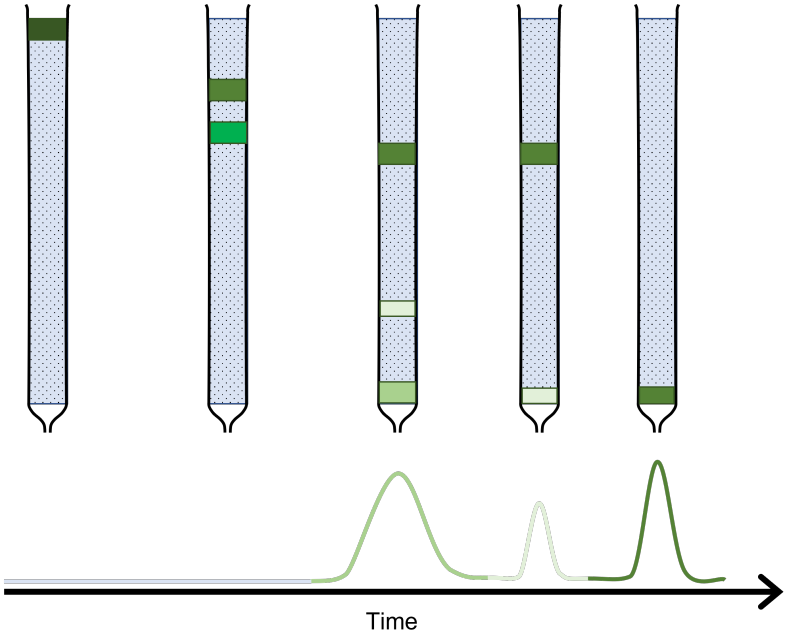

Figure 2.1 – Separation in chromatography. The dotted area represents the stationary phase. At the beginning of the elution, all the analytes are positioned in the same band. During the elution through the column, they move with different speeds, separating from each other, and will therefore have different retention times. As long as no analyte elutes, the signal measured by the detector is a flat line resulting from the constant flow of the mobile phase. When a molecule elutes and reaches the detector, the corresponding signal is recorded.

different compounds. Movement of a molecule through the column is only possible in the mobile phase, so the time required to pass through the column is related to the amount of time spent in this phase. Different molecules have different affinities for the mobile and stationary phase and therefore different velocities during elution; this is the key factor that separates molecules in chromatography. During the elution, each molecule is focused in specific bands; the final separation is achieved at the end of the column where the different bands can be collected or measured with a detector. Monitoring the elution process with a detector, it is possible to display the signal changes during the process as a function of time; the resulting plot is called chromatogram. In general a chromatogram is made of a flat constant signal, given by the continuous flow of the eluent, and peaks. The peaks represent the elution of a specific band through the end of the column, of course it means that the detector can interact somehow with the eluting molecule. In figure 2.1 a graphical representation of the process is given.

The chromatogram gives both qualitative (the presence or the absence of a given

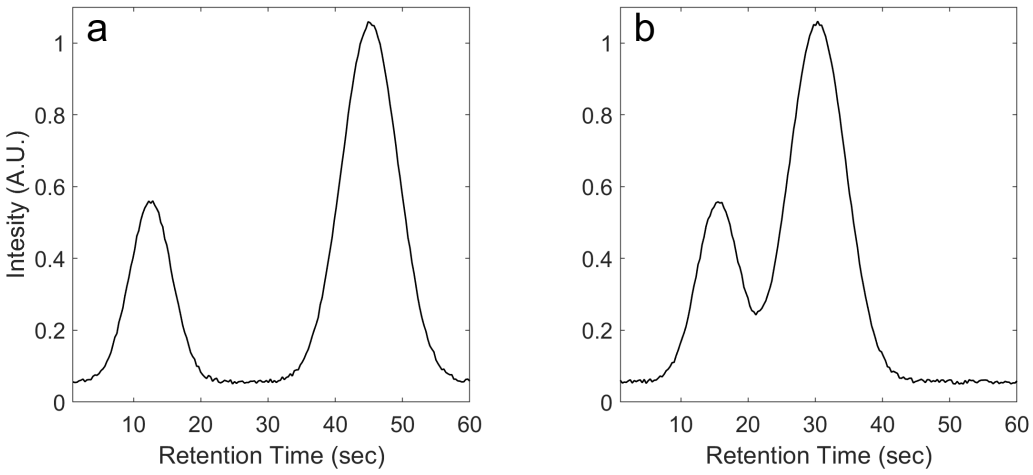

Figure 2.2 – Two chromatograms are represented. On the x axes is reported the Retention time in seconds, on the y axes the intensity of the signal in Arbitrary units (A.U.). a) Two resolved peaks. The elution of the compounds happens at different time points, thus the signals are not overlapping each other, this is an example of optimal resolution. b) Two unresolved peaks. The second peak starts to elute while the first is still coming out from the column. In this case the signals are overlapped and the peaks are not resolved.

molecule) and quantitative (the relative or absolute amount of each species) information. Ideally speaking, the peaks should be separated from each other in the chromatogram, meaning that the different compounds elute at different times. The more two adjacent peaks are distant to each other, the better they are separated. This can be measured in terms of resolution, which is the difference of the retention times of the two peaks divided by their avarage peak width. The retention time is defined as the time elapsed between sample introduction (beginning of the chromatogram) and the maximum signal (top of the peak) of the given compound at the detector, in other words it is the time required for a molecule to pass through the column (Skoog et al., 2013). In figure 2.2 an example of resolved and unresolved peaks is shown.

The resolution is determined by a number of different factors such as:

Total elution time The total time of the chromatographic experiment.

Mobile phase rate The flow rate of the mobile phase.

Distribution constants The ratio of the molar concentrations for a specific species in the stationary phase and in the mobile phase. It indicates the affinity of a specific molecule for the stationary phase.

Retention time The time required to a molecule to reach the detector. In particular it corresponds to the maximum of the corresponding peak in the chromatogram.

Selectivity factor The ratio of the distribution constants for 2 molecules, the higher is the ratio the more resolved are the corresponding peaks.

Number of plates The measure of the efficiency of the chromatographic system.

There are several rules and tricks to maximise the efficiency of a chromatographic run, i.e., to maximise the number of resolved peaks, nonetheless it is clear that at the increasing of the number of different molecules species in the analyzed sample, it becomes difficult to set experimental parameters that allow an optimal resolution for all the compounds. Moreover, some strategies can increase the resolution of the peaks but, at the same time, lead to other problems. For instance, a program of temperature ramp can increase the resolution of the peaks, but it can also cause drifts in the baseline signals.

2.3 Gas chromatography

Gas Chromatography (GC) and Liquid Chromatography (LC) are both widely applied in metabolomics (Beale et al., 2018; Zhou et al., 2012). The two approaches have a different, although overlapped, coverage of compound classes. The two main molecular features that impact the applicability of the two approaches are hydrophobicity and volatility (Brack et al., 2016). In general LC has a broader range of detectable species, but is characterized by a lower reproducibility and also the quantification of the samples is less accurate compared to GC (Lisec et al., 2006). Nonetheless, various classes of compounds, such as amino acids, carbohydrates, fatty acids, small molecular organic acids, phenolics, terpenes and sterols can be effectively separated with GC.

In GC the mobile phase is an inert gas, typically helium, argon or hydrogen. The samples are heated until the molecules reach the boiling point, at this point the molecules are injected in the column through a valve that mixes the vaporized sample and the mobile phase. The separation of the species follows the principles described

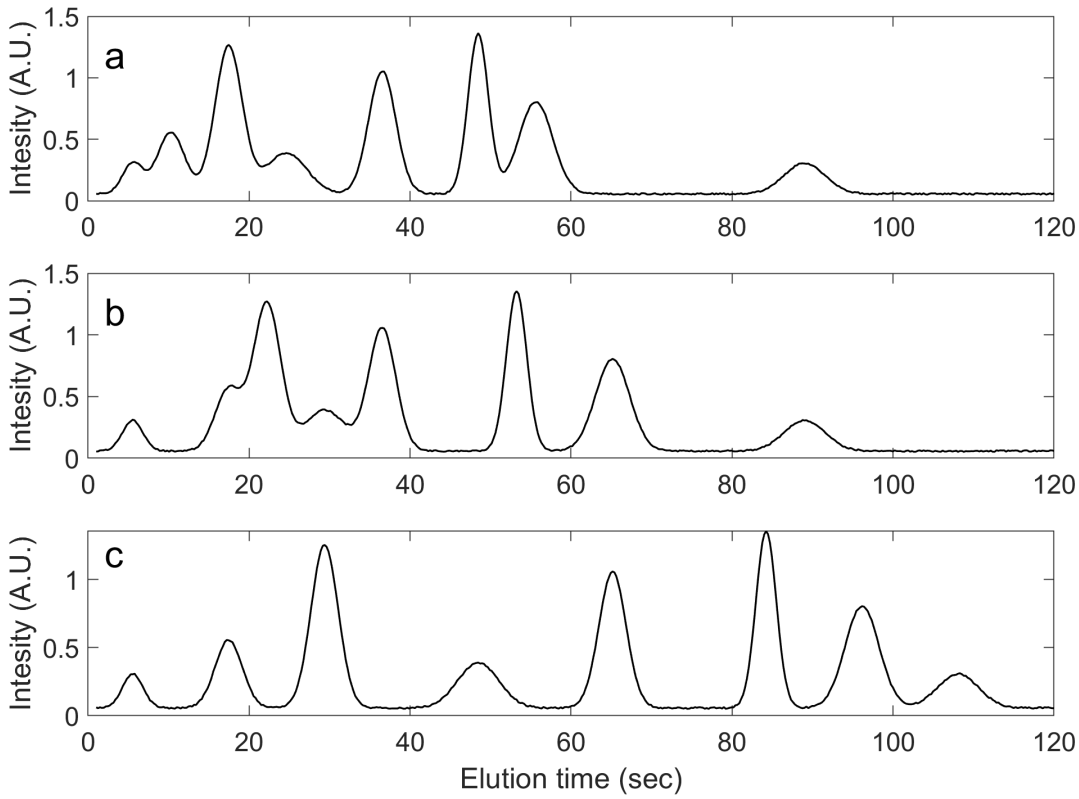

Figure 2.3 – Three chromatograms are represented. On the x axes is reported the chromatographic time in seconds, on the y axes the intensity of the signal in Arbitrary units (A.U.). a) Resulting chromatogram from an isothermal chromatographic run at low temperature. b) Resulting chromatogram from an isothermal chromatographic run at high temperature. c) Resulting chromatogram from a chromatographic run with a temperature gradient. It is possible to notice the improvements in terms of resolution of the peaks

in the previous paragraph. Given the complexity of the samples in terms of number of compounds in a metabolic profiling experiment, usually a ramp of temperature is applied to the column during the chromatographic run, which improves the efficiency of the chromatographic experiment (fig 2.3).

As mentioned in the previous section, chromatography can be coupled to a detector. In GC metabolic profiling the most common is a mass spectrometer.

2.4 Mass spectrometry

Mass spectrometry is a powerful analytical technique widely applied for the identification of molecules. In mass spectrometry (MS) the analyzed molecules are transformed into ions by applying energy. The ions are separated on the basis of the mass-to-charge ratio (m/z).

The main components of a mass spectrometer are shown in figure 2.4. In GC-MS

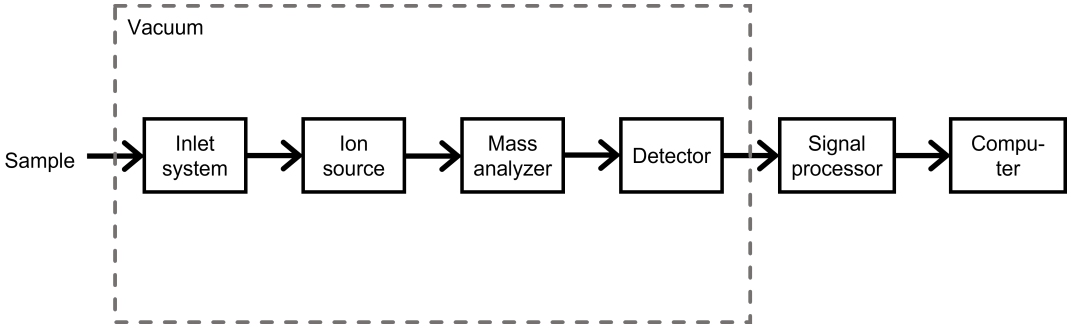

Figure 2.4 – Components of a mass spectrometer. The inlet system injects the sample into the instrument. The ion source ionises the molecules. The mass analyzer filters ions according to m/z. The detector counts the number of ions with the selected m/z. Finally the signal is processed and transferred to the computer.

devices the gold standard ion source is electron ionization (EI). With EI, the analytes are bombarded with a beam of electrons. The energy of the beam can remove an electron from the analyte or also fragment the molecules, obtaining a specific pattern of ions for each compound.

The ions are passed to a detector by a mass analyzer where the different ions are focused as a function of the specific m/z. The detector measures the abundance of the focused ions. Plotting the ions abundance versus the mass-to-charge ratios, it is possible to represent the specific ion pattern for a given molecule, called a mass spectrum.

Many different types of mass analyzers are available: Quadrupole, Time of Flight, Quadrupole Ion Trap, Ion Cyclotron. Quadrupole and Time of Flight are the most common mass analyzer coupled with GC systems (Lisec et al., 2006). In particular the quadrupole enables two modes of analysis: the full scan and the selected ion monitoring (SIM). In the full scan mode, a range of m/z ratios, typically from 50 to 500 m/z, is measured. In SIM only a set of specific m/z is recorded. While the first approach allows detection of a number of different compounds, the SIM mode is applied for the detection of a specific compound of interest. Intuitively, in metabolic profiling and metabolomics experiments, where the aim is to identify as much compounds as possible, the full scan mode is the most applied.

2.5 Gas chromatography - mass spectrometry

Nowadays more than 50 companies offer GC-MS platforms, suggesting how widely applied is this technique (Skoog et al., 2013). The two instruments are coupled on-line, thus the eluent from the GC is directly fed into the ion source of the mass spectrometer.

When the full scan mode is selected, at each scan the mass spectrometer measures the mass spectrum at the specified m/z range. Here the scan rate is crucial; a low scan rate increases the signal-to-noise ratio of the measurement, but the peaks could pass through undetected, a fast scan rate decreases signal-to-noise ratio, but more compounds are identified. Usually 2 to 5 scans per second is a good compromise.

Measuring the mass spectra as a function of time leads to a huge amount of data for each analyzed sample. In a common GC-MS analysis, the elution time is about 30 to 60 minutes. In a 60 minutes run, with a rate scan of 3 spectra/second, 12000 mass spectra are measured for each sample.

The eluting compounds are ionized with a high energy electrons beam, typically with a kinetic energy of 70 eV. This value is a convention in GC-MS analysis, but it gives some advantages. First the specific pattern of the ion fragments for a given molecule is reproducible across different instruments (Sparkman, Penton, and Kitson, 2011). This is not true with other ionization approaches such as electronspray ionisation, used in LC-MS, performed at lower energies and not totally consistent across different instruments.

Another advantage of EI is determined by the fact that, given the wide prevalence of this energy of ionization, the spectral libraries include spectra obtained at this energy. The spectral libraries play a fundamental role for the identification of the compounds in GC-MS analysis. Once a spectrum is obtained the corresponding compound is identified by comparing the experimental spectra with those included in the library. One of the most known libraries for EI-MS spectra is the NIST library and the last released version consists of 350,643 spectra of 306,869 unique compounds; most of

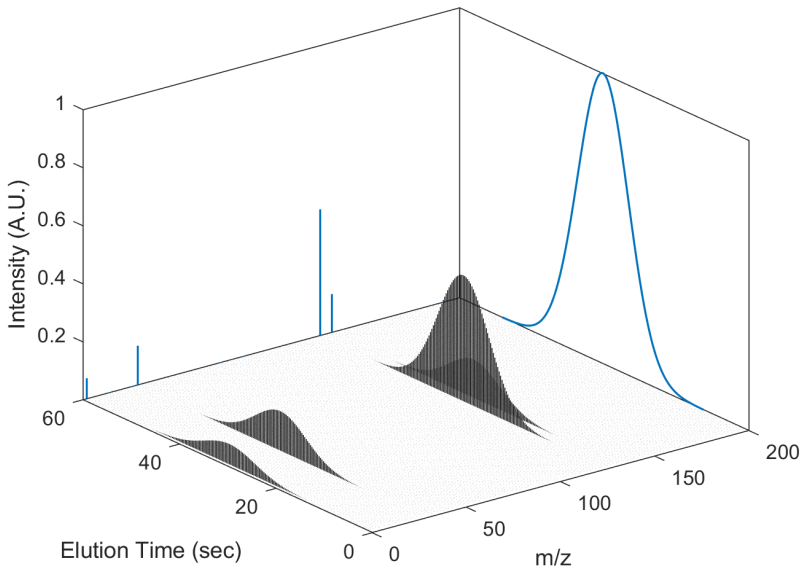

Figure 2.5 – Resulting surface of a GC-MS experiment. The mass spectra are measured as a function of time and can be displayed as a 3D surface (black areas). Another option is to sum the ion abundance for each scan resulting in the TIC (blue curve). The same operation in the opposite direction results in a summary of the mass spectra for a specific compound (blue lines).

which are measured at 70 eV (NIST, 2020).

The data from each GC-MS experiment are organized in a matrix where the rows correspond to the elution times and the columns to the measured m/z. A common way to display the data is the Total Ion Chromatogram or Total Ion Count (TIC), where the ion abundance of each spectrum is summed, or selecting a specific m/z resulting in a Extracted Ion Chromatogram (EIC) and, in both cases the result is similar to a conventional chromatogram, otherwise is possible to display all the measurements. In this case the result is a surface; an example is given in figure 2.5.

2.6 Quantification, identification, signals and artifacts

GC-MS data can be used for the quantification and identification of compounds in a mixture of unknown components.

2.6.1 Quantification

The quantification of a given compound can be absolute or relative. In order to obtain an absolute quantification of a given compound, it is necessary to build a calibration curve with the corresponding standard. In metabolomics the absolute quantification of compounds is possible with a targeted approach (Kapoore and Vaidyanathan, 2016; Røst et al., 2020), but in the untargeted approach the main interest are the relative concentrations for a given compound across the samples.

Theoretically, the quantification of a given compound across different samples is based on the comparison of the heights or the areas of the peaks related to the same compound. When a single sample is analyzed, both the height and area can be used for the quantification, but usually the estimation of the height is preferred over the area since it is easier to obtain.

Instead, when multiple samples are analyzed the area gives a more robust estimation. In untargeted GC-MS experiments it is common to observe changes in the peaks, in terms of width and shape, from sample to sample. This means that the height of the peaks can determine inaccurate quantification as shown in figure 2.6.

However, the estimation of the area under the peaks presents different issues. The first one is that the signal measured by the detector is made of different components: the signal from the baseline, the signal from the compound, and the experimental noise (Amigo, Skov, and Bro, 2010) (figure 2.7). The signal from the baseline consists in a flat constant line, even though drifts in the signals are common, due to the temperature ramp during the chromatographic run. The compound signal instead has specific boundaries, defined by the time needed for the total elution. The noise signal, a random variation mainly dependent on the detector and also on the sample matrix, is constant throughout the signal detection. Thus, in order to quantify the area under the peaks, two ways are possible: extract the specific signal of the peak or integrate the whole area. The first case results in more accurate quantification, but is not always clear how to obtain the original signal. In the second case, it is required to set arbitrary boundaries, since the beginning and the end of the peak are hidden

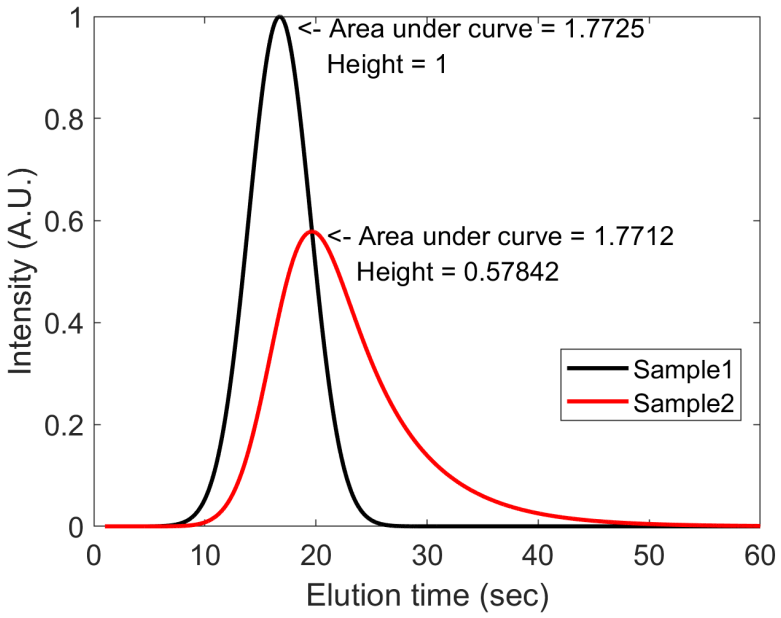

Figure 2.6 – A graphical example to show the different results when the height or the area are estimated for the quantification of a peak. In black and red are represented two chromatograms from two different samples for the same compound. The molecule has the same quantity in the two samples. On the x axis is reported the elution time, on the y axis the intensity of the signal. Assuming that the signal increases linearly with the number of molecules passing through the detector, the two curves can be directly compared. When the height is used to quantify the compounds, it results that the quantity of the compound in sample 1 is almost doubled compared to sample 2. Instead integrating the area under the two peaks the estimation is correct

by the total signal. The definition of the boundaries and the resulting width of the interval can dramatically impact the final estimation as shown in figure 2.7.

Moreover, it is likely that different chromatographic runs will have different intensity of the baseline, which intuitively makes even more difficult an accurate estimation of the same compound across different samples.

When two signals overlap each other, there are different ways to estimated the limits of the peaks. The most common approaches, depicted in figure 2.8 are:

Drop method A straight vertical line is drawn starting from the valley point between the two peaks. The estimations are not accurate, in particular for the smaller peak. The error depends on many factors such as the ratio of the dimensions of the two peaks, the degree of overlap and the asymmetry. Only in specific (unlikely) cases, the estimations are correct and the following conditions must be met (Dyson and Smith, 1998):

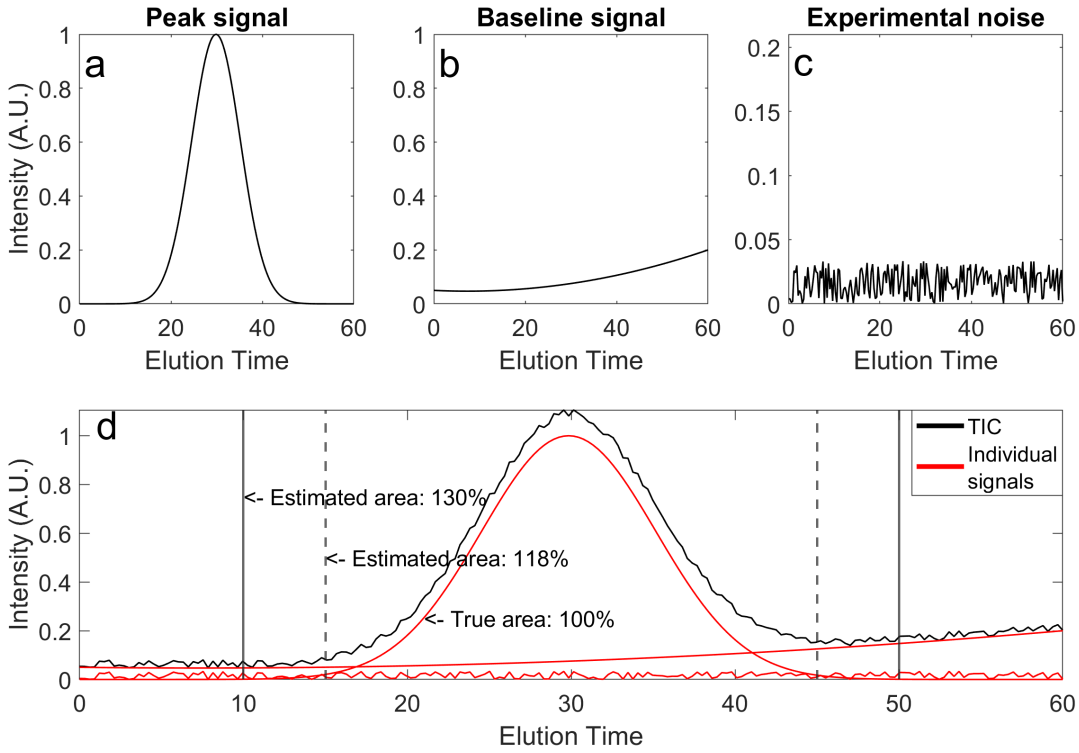

Figure 2.7 –
a, b and c: Individual contributions in a chromatographic signal
d: Resulting estimated area within different boundaries. Two sets of boundaries have been defined (solid black lines and dotted black lines). The corresponding estimated area, considering the TIC, are reported. The original area for the red peak is reported as 100%, the areas evaluated directly from the TIC are reported as the difference to the true area. Notice that the wider interval, defined by the black lines, leads to an over estimation of 30% compared to the true area. Instead with a narrower interval the overestimation is greatly reduced.

1. the peaks are equal;
2. if the peaks have different heights, the valley is less then 5% of the smallest peak;
3. the baseline is flat
4. the peak limits and the valley point can be precisely identified.

Valley method The start and stop points are set at the valley between the peaks. For each peak a straight line is drawn to the left or the right connected with the baseline. This method results in underestimation for both the peaks, particularly for the smaller peak.

Exponential skim method An exponential curve is fitted starting from the valley point to the baseline for each peak. Here shoulder peaks, i.e., peaks emerging on

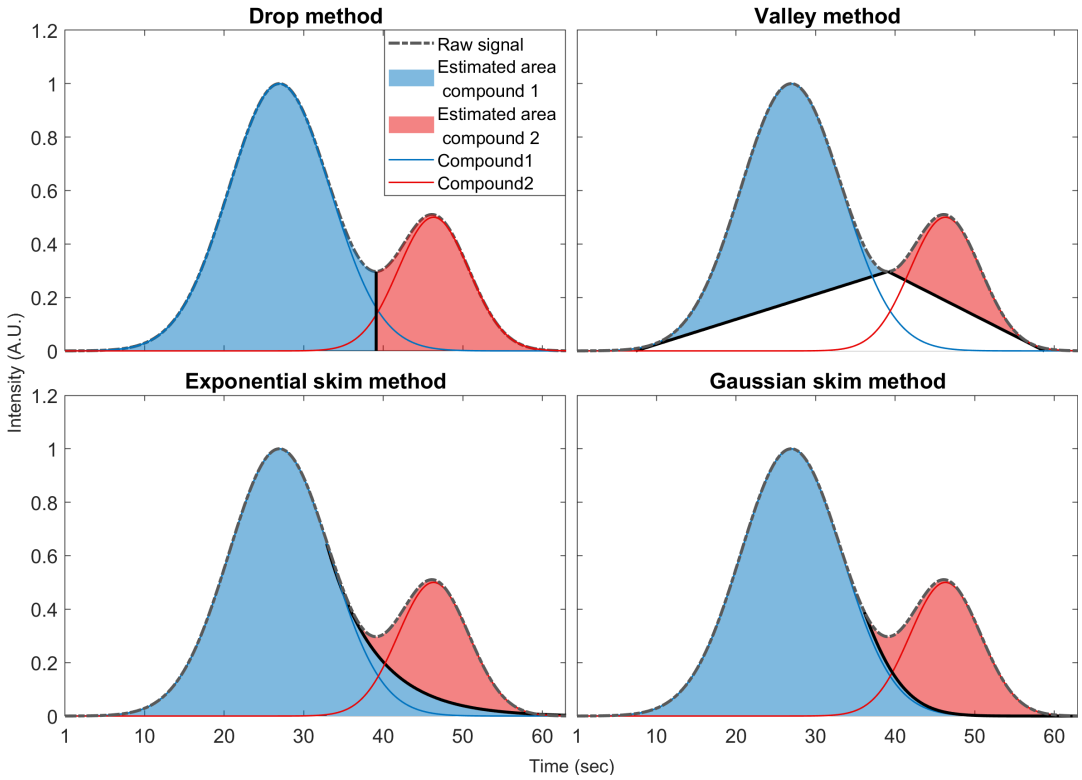

Figure 2.8 – Graphical representation of common methods for the estimation of the area under coeluting peaks.

the tails of other peaks, are underestimated (Bicking, 2006).

Gaussian skim method Similar to the exponential skim method but a Gaussian curve is fitted. This method assumes Gaussian shapes for chromatographic peaks, but, although this is a nice approximation, this is not the case for many peaks.

In addition, all these methods (graphically represented in figure 2.8) require a valley point, so it must be evident that two compounds are eluting at the same time. However, in severe co-elution, different compounds can elute at the same time, making the detection difficult and these methods inapplicable as shown in figure 2.9. It is still controversial whether the height or the area should be used to quantify the peaks (Dyson and Smith, 1998; Bicking, 2006). However in the cases of absolute quantification and semi relative quantification across different samples the estimation of the area is more robust compared to height (Dyson and Smith, 1998).

2.6.2 Identification

In untargeted analysis, the identification of the compounds relies on the comparison of the experimental (measured) mass spectra with respect to spectral libraries. The main distance applied to compare different spectra is the cosine distance (Stein and Scott, 1994; Stein, 1999), defined as:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \sqrt{\sum_{i=1}^{n} (y_i)^2}} \tag{2.1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the two vectors representing the mass spectra and $\mathbf{n}$ correspond to the number of m/z values measured by the instrument. The cosine distance measures the angle between two vectors, so if they are pointing the same direction the angle is 0° and the cosine is 1.

The mass spectra can be directly extracted by the experimental data. Usually the mass spectra at the maximum of the peak is considered. However, the extracted spectra will include the signals from the compounds, the baseline, eventually co eluting compounds, strongly limiting a correct identification . A common preprocessing for the spectra identification is to subtract the signal of the baseline taking the spectra in a flat region of the chromatogram (Want and Masson, 2011). However, also in this case, the coelution of the compounds is a major obstacle to the accurate identification of the eluting compounds. If the overlap of different compounds is not evident, the spectra resulting from the sum of the compounds will be used during the comparison with the spectral library, determining incorrect identification; an example is given in figure 2.9.

2.7 Conclusions

GC-MS is a powerful analytical technique, widely applied in metabolomics experiments. The combination of chromatography with mass spectrometry allows the detection of a number of compounds. Nonetheless, given the complexity of the analyzed samples, the number of molecules and the sum of different signals, the appropriate quantification and identification of the compounds can be difficult. From the few examples shown in this chapter, it is evident that the extraction of the relevant information can be

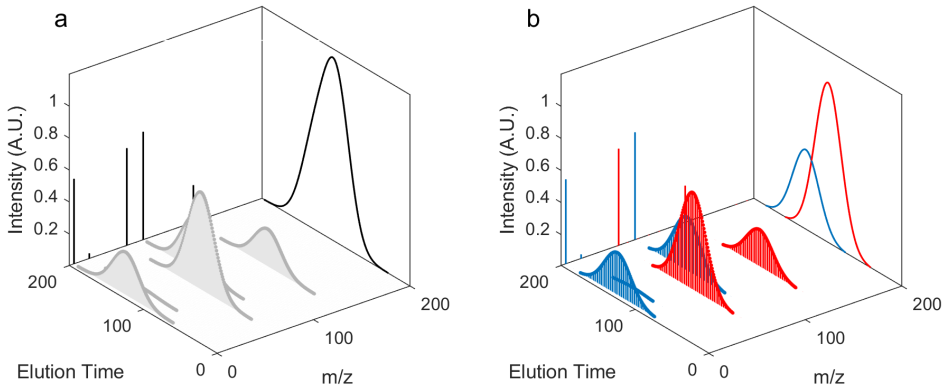

Figure 2.9 –
a: Total signal from the GC-MS experiment. The TIC (black curves), the surface (grey surface), and the m/z (black lines) are shown
b: The same signal is decomposed in the individual contributions showing that two compounds are eluting at the same time.
The inspection of the TIC does not reveal the presence of two compounds (blue and red). Using the raw data will lead to incorrect identification and quantification

difficult to reproduce, time consuming and requires specific know-how. In the next chapter an overview of the software developed specifically for this task will be given.

Chapter 3

Tools for the analysis of GC-MS data

As described in the previous chapter, the identification and quantification of the chemical compounds from GC-MS experiments is a difficult task to handle. This is due to the huge number of chemical compounds in the samples, the different interfering signals and the difficulty to optimize the experimental settings to obtain resolved peaks.

In order to assist researchers in the identification and quantification of the compounds from GC-MS data, several tools have been developed. In this chapter an overview of the most used software in metabolomics is given. Some of these methods have been developed for LC-MS analysis, nonetheless they are also applied for GC-MS data.

All the software described here are freely available, since details about the functioning of commercial tools, such as ChromaTOF (LECO corporation), Compound Discoverer and Chromeleon (Thermo Fisher Scientific), MassHunter and ChemStation (Agilent), are not shared by the vendors.

3.1 MZmine

Katajamaa and Orešič proposed MZmine for the first time in 2005 (Katajamaa and Orešič, 2005), a major update of the platform has been released in 2010 (Pluskal et al., 2010). The original focus of the platform was the analysis of LC-MS data, nonetheless

the approach has been widely applied also for the analysis of GC-MS datasets (Pinto et al., 2018; Araújo et al., 2018; Huang et al., 2013).

The software has been developed in java and a standalone Graphical User Interface (GUI) is available to help users during the analysis. The overall workflow for the identification and quantification of the peaks is based on a sequence of steps.

Software architecture MZmine is a collection of methods for the analysis of LC or GC-MS data, the methods are organized in different independent modules:

- data import;
- data preprocessing;
- visualization;
- peak detection;
- peak identification;

each module includes the respective approaches.

Data import In the first version the only file format accepted was .cdf, while MZmine2 accepts the following file formats :

- .cdf
- .mzML
- .mxXML
- Thermo Fisher raw format

Data visualisation To visualize the loaded data, different plots are available: TIC, 2D, 3D. Another option to visualise the data is to consider only the most intense m/z for each scan, the result is similar to the TIC and it is called Base Peak Chromatogram (BPC). The mass spectra can be visualized in a continuous mode, where the actual m/z are plotted or with a centroided mode (this is mostly related to the analysis of LC-MS data where centroiding/binning of the adjacent m/z is a common solution to handle high resolution mass spectra).

1- Peak detection In MZmine the quantification and identification of the peaks is called *peak detection*. The peak detection is a three steps process and is performed sample wise, meaning that each sample is analyzed independently from the others. The analysis can be performed on batch of data, thus different samples are analyzed at the same time with the same settings. The three steps are:

2- Mass detection The mass detection step generates a list of selected m/z ratios for each mass spectrum. The software includes five different algorithms to perform this step: centroid, exact mass, local maxima, recursive threshold and wavelet transform. Each of the methods requires the user to set thresholds, for instance the centroid method removes all the m/z signals below a given threshold set by the user. Moreover, the user should be aware of which method is the best option for his/her own data. The result is a list of candidate m/z ratios to represent a given compound. A list is obtained for each mass spectrum obtained for each of the analyzed samples.

3- Chromatogram builder In the second step the chromatogram builder uses the mass list generated during the mass detection step to connect m/z spanning over multiple scans. Basically it defines EICs only on selected masses. To this end the method suggested by the authors is the Automated Data Analysis Pipeline (ADAP) algorithm (Smirnov et al., 2018), recently implemented in the MZmine platform. First all the selected m/z for a given sample are sorted by their intensity. The first EIC is created according to the most intense m/z. Then the next m/z from the sorted list is considered. If it is the same m/z to a previously EIC defined then is added to the given EIC, otherwise a new EIC is created. The process loops over all the m/z extracted. At the end a list of EICs is obtained. However, not all the EICs are considered. During this step four parameters must be defined by the user in order to select which EICs will be considered for the next steps:

- Group intensity threshold, an intensity threshold.

- Minimum group size in number of scans, the minimum number of sequential scans in a given EIC having points above the group intensity threshold set by the user.
- Minimum highest intensity, an intensity threshold. At least one m/z intensity in the EIC must be higher or equal to the threshold.
- m/z tolerance. Maximum m/z difference of data points in consecutive scans in order to be connected to the same EIC.

The idea is that each ion is specific for a given compound so a single peak should be reconstructed for each chromatogram. However, it is common that more peaks are included in a given EIC. The next step aims to resolve this issue.

Chromatogram deconvolution In this step the reconstructed chromatograms are deconvolved into individual peaks in order obtain single peaks for each EIC obtained by the previous step. The peak deconvolution is performed by means of filters, in particular the continuous wavelet transform using Mexican hat wavelet is applied. Basically it applies filter with increasing width, looking for the regions of the chromatogram where there are peak(s).

Six parameters must be defined by the user for this step:

- a threshold to define the signal-to-noyse ratio, to remove noise peaks;
- a threshold to define the minimum intensity a peak can have to be considered as a real peak;
- a threshold to define the minimum area under the peak;
- the minimum and maximum width of the peaks;
- the range of widths for the wavelet filter.

A given region can include one or more peaks and each of them is analyzed to assess the number of putative compounds included. At this point a double round hierarchical clustering to deconvolve the actual spectrum of the compound is performed. In the first round the identified region are clustered together accordingly to the retention time. At this point the number of compounds described by the region is estimated by means of five different criteria. For each region five

metrics are calculated and a score is assigned to assess both if a a given region is actually representing an eluting compound and also the number of peaks, i.e., the number of compounds, included.

The deconvolution of the spectra is based on the detection of coherent trend of the m/z in each region. To do so a second hierarchical clustering approach is used to group together the m/z with similar patterns within a given region, thus discriminating between different compounds that should have specific ion fragments. The number of clusters is defined accordingly to the number of compounds detected for a given region. The second clustering round requires two settings to be defined by the user:

- the minimum distance between the analytes in terms of elution time;
- a threshold to set the minimum difference between peaks, defined as the cosine distance of the reconstructed elution profiles.

Peak identification The identification is done by comparing the masses of the deconvolved peaks with internal or external libraries. This step is performed for each peak detected in each sample. Thus, after the identification, the final step consists in the alignment of the peaks detected for each sample in order to identify all the peaks referring to the same compound across the different samples.

MZmine2 also includes some tools for the analysis of the data such as PCA, clustering and heat maps. It is interesting to notice that when PCA is performed the user is not required to set anything about preprocessing and scaling of the data. This have been extensively showed as a critical point for the analysis of metabolomics data (Berg et al., 2006).

3.2 AMDIS

AMDIS is an acronym for Automated Mass spectral Deconvolution and Identification System. It is a freely available software with a dedicated GUI. It has been developed at NIST (National Institute of Standards and Technology), a US government agency for

the detection of chemicals in violation of chemical weapons conventions. The main core of this approach is the deconvolution method for the analysis of mass spectra originally proposed by Stein in 1999 (Stein, 1999). The algorithm involves four sequential steps:

- noise analysis;
- component perception;
- spectral deconvolution
- compound identification.

Only one file at a time can be analyzed and no batch analysis can be performed, meaning that only a sample at a time can be analyzed.

Data import Different formats can be loaded:

- Bruker
- Finnigan (GCQ, INCOS, and ITDS formats)
- HP
- MassLynx NT
- MicroMass
- Perkin-Elmer
- Shimadzu MS Files
- Varian Saturn Files
- Viking

Data Visualization Once the sample is loaded, the plot of the TIC is visualized in the main window and selecting a specific scan is possible to look at the corresponding mass spectrum.

Settings Before resolving the different mass spectra the user has to define the settings for the algorithm. Default settings are available for the automatic analysis, but a careful evaluation is suggested in the AMDIS manual to obtain better results (D'Arcy and Mallard, 2004). A number of parameters can be adjusted.

Component width A threshold to define the maximum width of a peak.

Omit m/z This option allows to automatically exclude specific m/z ratio(s). Any m/z values specified will not be considered for defining what constitutes a chromatographic peak. However, m/z values will be used when the analysis process extracts a spectrum and so the m/z peak may end up in the extracted spectrum

Adjacent peak subtraction It can happen that adjacent m/z ratios in a given mass spectra are related to different compounds. Setting this option to one or two allows the software to exclude one or two interfering ions. To enable the adjacent peak subtraction is advisable when the mass spectra are crowded with signals which could suggest the presence of coeluting compounds. However, the system could be able to find coeluting compounds even without the Adjacent peak subtraction.

Resolution Three resolution values can be selected: high, medium, low. The higher is the resolution the more the algorithm separates adjacent peaks, increasing the number of compounds detected.

Sensitivity Five sensitivity values can be selected: very high, high, medium, low and very low. The higher is the sensitivity the more the system considers the signals from the scans as actual signals instead of noise. For metabolomics experiment it is advisable to keep it low considering the sub optimal experimental settings and the complexity of the samples.

Shape requirements This option sets how near the shape of the TIC has to be similar to a peak to actually consider a given profile as a resolved component. Three options are available: high, medium, low, the higher is the requirement the more strict is the requirement the resolved peaks have a peak shape.

Another set of parameters are called filters. These are categorical information that the system considers during the analysis of the spectra. Examples are: the minimum number of m/z to be included in a peak, the minimum threshold for the signal to noise ratio, the minimum abundance to consider a signal as a peak. All these parameters have to be manually selected and small differences in the settings can dramatically change the result of the analysis.

Peak identification As mentioned above the algorithm involves 4 steps.

Noise analysis In the first step a noise analysis is performed in order to extract:

- Noise factor, an estimate of the overall intensity of the noise for the entire chromatogram
- Threshold transition, an estimation of the values for the measurement under the limit of detection, arbitrarily set to 0.
- m/z Peak Uniqueness, consisting in the evaluation of the uniqueness of each m/z.

Component perception The next step consists of the component perception. A component, i.e. a compound, is perceived when a sufficient magnitude of its ions maximise together. To assess this trend a number of criteria are considered inspired by classic theory of chromatography analysis.

Spectral deconvolution Finally, for each of the perceived components a mass spectrum is deconvolved. The method is based on the original approach based on least squares estimation proposed by Dromey and colleagues (Dromey et al., 1976). The intuition of the approach is to find the m/z that coherently fluctuate up and down and group them together to build the estimated mass spectrum. For the deconvolution, different thresholds need to be specified as well as assumptions on the baseline trend and the peak width, which depend on the parameters selected before the analysis

Identification The identification is finally obtained by comparing the deconvolved spectra of each identified peak with the NIST library, developed in the same institution. An estimation of the area under the peaks is obtained

by building the TIC only with the deconvolved spectra and integrating the underlying area.

3.3 MS-DIAL 2.0

MS-DIAL was originally proposed for the analysis of LC-MS-MS data (Tsugawa et al., 2015) but in the version 2.0 also the analysis of GC-MS data has been implemented (Lai et al., 2018). The software includes a GUI.
Different formats can be loaded:

3.3.1 Data import

- Bruker
- Sciex
- Waters
- Thermo Fischer
- Perkin-Elmer
- Shimadzu MS Files
- Agilent
- .cdf

3.3.2 Quantification and identification

The quantification and identification of the peaks involves the following steps:

Profile smoothing The first step consists in the smoothing of the chromatogram. Several approaches are available, among them the most recommended are Savitzky-Golay and linear weighted moving average. Savitzky-Golay is a widely applied smoothing approach, the smoothing is achieved by fitting adjacent data points with a polynomial using the linear least squares approach (Savitzky and Golay, 1964). The user has to set the parameters to perform the smoothing such as the width of the window for the polynomial.

Peak spotting The next step is called *Peak spotting*. Here the user needs to specify if the data are at high or nominal resolution. If the data are at high resolution, ranges of m/z are gathered and the Base Peak Chromatogram (BPC) is extracted, otherwise each m/z is considered obtaining standard EIC. The BPCs or the EICs are analyzed in order to detect the peaks. The detection of the peaks is based on the fitting of a Gaussian shape curves. The fit of the curve to the BPC or EIC is evaluated using different parameters with specific thresholds, such as peak quality, peak sharpness and threshold that have been defined to discriminate between good and bad candidate peaks. Subsequently the peaks (called peak spots by the authors) derived from different m/z ratios with the same width and the same retention time (but tolerance thresholds for retention time shifts can be set) are grouped together.

Peak deconvolution For each spotted peak a deconvolution step is performed to assign the corresponding mass spectra. The deconvolution is based on a least squares regression model based on the identified peaks. The purpose of the deconvolution is to estimate the specific peak abundance of m/z traces shared among coeluting metabolites. This is done through the definition of model peak m/z traces in the retention time region of each peak group. The deconvolution allows the detection of the unique spectra of coeluting compounds with a maximum of four compounds

Peak identification The identification of the peaks can be performed by comparing the deconvolved spectra with libraries; in the original publication the NIST library is suggested.

3.4 eRah

eRah is an R package for the quantification and annotation of GC-MS datasets (Domingo-Almenara et al., 2016). The software does not include a GUI, thus some knowledge of the R suite is required to use it.

The accepted format for the data is .mzXML. The identification of the peaks involves five steps:

Preprocessing The denoising of the raw data is performed using the Savitzky-Golay method in order to smooth the profiles and filter the baseline signal. The width of the window for the calculation of the polynomial must be specified by the user.

Spectral deconvolution The deconvolution is a two step process. First a multivariate filter called Compound Match by Local Covariance (CMLC) is applied to each sample matrix and detects the regions of the chromatogram where the compounds are eluting. The user has to set the minimum peak width, the noise threshold and also specific m/z can be excluded. It is worth noting that the filter looks for regions where the signal has a Gaussian shape, thus not optimally eluting regions might not be detected. In the second step, once the peaks have been detected, a deconvolution algorithm is applied to reconstruct the specific mass spectra. Specifically, the Orthogonal Signal Deconvolution (OSD) is applied . Intuitively, this method finds the ions that are covaring in the regions detected in the first step. Finally, the elution profile corresponding to the reconstructed spectra is calculated. This is done with a least absolute deviation (LAD) regression (Li and Arce, 2004) between the spectrum found by OSD-resolved spectra and the chromatogram. These steps are performed for each sample and for each region of interest.

Alignment The deconvolution is performed independently for each sample at a time, thus to compare different samples it is needed to identify the peaks that are describing the same molecule across different samples. Thus, the next step is the alignment of the deconvolved profiles. This is a local alignment approach; the algorithm searches the peaks deconvolved in the different samples within defined windows on the retention time dimension and groups the peaks together if they are detected as describing the same chemical. Here the user has to set the maximum allowed time shift and the minimum spectral similarity and both these parameters can strongly affect the output of this step.

Compound identification In the last step, the aligned peaks are identified by comparing the mean spectra for a given compound with a spectral library. At

the same time, an estimation of the relative concentration across the sample for each compound is calculated by using the deconvolved profiles.

3.5 XCMS

XCMS is an R package proposed in 2006 (Smith et al., 2006). Originally the knowledge of the R suite was required in order to use the platform, but an online version of XCMS has been released in 2012 with a dedicated GUI, resolving this obstacle (Tautenhahn et al., 2012). Originally proposed for LC-MS analysis, nowadays it is also applied for the analysis of GC-MS data (Arbona et al., 2009; Liu et al., 2014; Ueberschaar et al., 2021).

Different data formats can be loaded:

3.5.1 Data import

- mzData
- .mzXML
- Agilent
- CDF

3.5.2 Quantification and identification

The quantification and identification of the peaks is performed sample wise, but batch analysis is possible. The following steps are involved:

Peak detection XCMS works on one sample at a time. The m/z values are grouped (the span of m/z for the groups is defined by the user) and for each group the highest ion intensity for each scan is extracted. The result of this step is called Extracted Ion Base Peak Chromatogram (EIBPC). For each EIBPC, a filter is applied to detect the presence of peaks. The filter has the shape of a second derivative Gaussian curve, similar to the filter applied in eRah. The width of the Gaussian has to be defined by the user, as well as a signal-to noise ratio threshold. Only the peaks above this threshold are actually considered as peaks. The noise is defined as the mean signal from the unfiltered data.

Peak alignment All the peaks detected in all the samples have to be aligned in order to compare the differences across the different experiments. This is done by grouping together all the detected peaks, then the system identifies peaks from different samples that occur at the same retention time. The alignment is based on two steps. In the first step, an algorithm called PeakDensity, based on another filter, finds groups of peaks, considering the retention time, across samples, already aligned. Also in this case the width of the filter must be set by the user. These groups, called Well Behaved Peaks Groups (WBPGs), are used as a first reference for the subsequent alignment. The median retention time and the deviation from the median for each sample is calculated for every group. Using the median retention time of each WBPG a local alignment is performed for each sample. The final output of this step is a list of features, i.e. set of peaks across samples for the same compound. Also here some thresholds are considered, for instance a peak is discarded if less then 50% of the samples show a given peak.

Peak Identification In the XCMS platform a dedicated method has been specifically developed for GC-MS untargeted data, called metaMS (Wehrens, Weingart, and Mattivi, 2014). The identification of a peak is based on the creation of a pseudo spectrum that is compared with available libraries. The pseudo spectra generation is based on the CAMERA package (Kuhl et al., 2012), and it is based on the clustering of the m/z signals for each of the regions identified in the peak detection phase. Different filters and thresholds can be applied during the construction of the pseudo spectra, such as the elimination of peaks with low intensity, or the elimination of pseudo spectra with few peaks.

3.6 MetAlign

The MetAlign software was proposed in 2009 (Lommen, 2009), and since then numerous updates have been released. The method was designed both for the analysis of LC and GC-MS data and, as the author declared, it is an expert system inspired by the manual analysis of the data performed by expert analysts. It includes a GUI.

Different data formats can be loaded:

- MassLynx
- Xcalibur
- mzData
- .mzXML
- Agilent
- CDF

The identification and the quantification of the compounds is performed with different steps:

Preprocessing In the first step, the baseline and the noise are estimated and removed from the data; nine options need to be selected by the user. The denoising of the data is based on a filter applied on the mass direction. The baseline correction is based on a local linear estimation.

Peak detection A peak is defined looking at the preprocessed mass spectra. If two adjacent scans in a mass trace increase more than a given threshold of local noise value, going from left to right or from right to left in the mass trace, the possibility of a mass peak occurring is noted. The software can add other possible adjacent points participating in the potential mass peak by loosening the local peak finding criteria (lowering the threshold). All retention times in a mass trace noted as potential peaks are placed in *peak regions*, while all points not noted are placed in *baseline regions*. The baseline regions are set to zero and the peak detection is performed for a second time, considering a half of the local noise value. A final selection of the peaks is performed in order to remove small (noise) peaks. The threshold for the final selection is based on the noise in the data and is defined by the user.

Alignment Finally the peaks identified in different samples are aligned. Two algorithms are available called rough and iterative (suggested for crowded datasets). Different options must be specified by the user also in this case. The aligned data are used for the final identification and quantification of the compounds.

3.7 Tools comparison

In the above paragraphs, some of the most common software for the analysis of GC-MS data have been described. However, several tools are proposed every year, and different publications are available for an updated review of many other platforms (Spicer et al., 2017; Misra, 2018; O'Shea and Misra, 2020; Misra, 2021).

Given the number of different options, an important point is which platform should be selected to perform the analysis of the data.

Myers et al. performed an accurate comparison of XCMS and MZmine2 analyzing a mixture of standards and human plasma samples (Myers et al., 2017). In the paper they showed a strong impact of the thresholds for the identification of the peaks for both the methods. Furthermore, despite the common features detected by both systems, specific peaks not observed by the other were obtained from each software. This implies either that the analysis of the data must be performed with more than one tool to detect all the compounds, or that different false positive and false negative peaks are detected by both systems. The authors provide a list of suggestions for users to minimize errors during dataset analysis.

Another comparison of software, particularly focused on the alignment of the peaks, has been performed by Koh and colleagues (Koh et al., 2010). This is a crucial step in most of the tools described above, since the correct alignment of the peaks is required for the comparison of different samples. Also in this case the authors pointed out the importance of the settings and the impact of several parameters for the good performances of the analysis. Moreover, they emphasized the risks of misleading results using tools designed for LC-MS with GC-MS data.

Coble and Fraga (Coble and Fraga, 2014) compared MetAlign, MZmine, SpectConnect, and XCMS for the analysis of both GC and LC-MS data. Also in this case great discrepancies among the results from the tested software have been highlighted. The authors suggested the use of multiple platforms for an accurate identification of the compounds from the experimental data. Moreover, they stressed the point of the large

Table 3.1 – List of publications comparing software for GC-MS data analysis

Publication	Compared software	Data
(Myers et al., 2017)	XCMS MZmine2 MetAlign	-Standards mixture - 21 compounds -Human plasma samples -Standards mixtures
(Koh et al., 2010)	MZmine TagFinder MetAlign	-Bladder cancer and non-bladder cancer urine samples
(Coble and Fraga, 2014)	MZmine SpectConnect XCMS	-Standards mixtures
(Li et al., 2018)	XCMS MZmine2 MS-Dial MarkerView Compound Discoverer	-Standards mixtures -Black pepper metabolite extracts
(Niu et al., 2014)	XCMS SpectConnect MetaboliteDetector MetAlign MZmine TagFinder MeltDB Gavin	-Standard mixtures

number of false positive peaks detected by all the systems, suggesting a thoroughly manual analysis of the results. Other comparison publications are listed in table 3.1.

The comparison of the available tools is definitely important for the researches looking for methods to analyse the data. Nonetheless, a couple of considerations are needed. The first is that the analysis of standards mixture is appealing since the results can be validated and the comparison gives more quantitative indications. However, it should be noted that the complexity of real data is incomparable with prepared samples, and nice results with standards mixture do not imply reliable results with real samples.

A second consideration is about the expertise required to use these software. The ability and the knowledge of the software is of primary importance for the selection of the right settings and is not always clear if the authors have the same skills for all the tested approaches. In this contest, an interesting emerging field is the development of meta models for the selection of the best parameters during the analysis (Lassen et al., 2021; Libiseller et al., 2015).

3.8 Conclusions

All efforts of the researchers for the development of new tools, the comparison of already existing platforms and the analysis of datasets performed with multiple solutions suggest that there is still room for improvement and that a general accordance regarding the best approach for the peak detection identification and quantification is still missing. At the same time the comparison of the existing tools confirms the need for better solutions for the analysis of GC-MS data. In particular, the most critical steps are:

- the definition of the settings required to perform the analysis of the data;
- the independent analysis of the samples, which are compared once the information has been extracted;

The approach proposed in this thesis, described in the next chapters, attempts to resolve these issues, giving a reliable and fully reproducible list of compounds.

Chapter 4

Methods

In this section the methods applied during this project are described. The section will describes with the definition of the notation and the description of bilinear and trilinear models. The second part describes two alignment approaches for the preprocessing of chromatographic data. In the last part artificial neural networks, for the classification of the peak shapes are introduced.

4.1 Notation and terminology

The following notation has been adopted:

- scalars are denoted by italics lowercase symbols. Example: x;
- one way matrices, i.e., vectors, are denoted by bold lowercase symbols. Example: $\mathbf{x}$;
- two way matrices are denoted by bold uppercase symbols. Example: $\mathbf{X}$;
- three and higher way arrays are denoted by underlined bold uppercase symbols. Example: $\underline{\mathbf{X}}$;
- the number of features of a matrix are denoted by:
  - I for the first dimension;
  - J for the second dimension;
  - K for the third dimension;

- A model is denoted by the symbol ˆ. For example the model of $\mathbf{X}$ is denoted by $\hat{\mathbf{X}}$;
- the number of components included in a model is denoted by: F;
- the matrix of residuals is denoted by: $\mathbf{E}$.

4.2 Models

Summarizing, the aim of this project was to develop an automatic approach for the quantification and identification of molecules from the signals measured by a GC-MS instrument.
As described in chapter 2, the raw signal measured in GC-MS experiments is actually the sum of different contributions and the focus is the quantification and identification of the compounds. Thus, a preliminary step is the estimation of the different sources and the identification of the interesting signals, i.e., the peaks.

The tools described in chapter 3 are based on extensive preprocessing and semi-manual analysis of the experimental data, in order to highlight the peaks that are describing chemical compounds. In short, the strategies can be described as a heuristic approach, combining different criteria and methods for the extraction of the molecular signals.

Another possible way to extract the single contributions from complex mixtures is to build a model to describe the data. In this sense a model is defined as an approximation of the data (Bro, 1998). In general a model is defined by:

- structure
- parameters;
- constraint(s);
- loss function.

The core of the AutoDise algorithm, developed in this thesis, is based on a model called PARAFAC2. In the first part of the following section the concept of model and

some of the related aspects are described, in the second part a brief introduction to the theory behind deep neural networks is given. In the last part, the classification metrics applied during this project are introduced.

4.2.1 PCA

A Principal Component Analysis (PCA) model can be used as an example to introduce the concept of model. A matrix $\mathbf{X}$ with dimensions IxJ is approximated by a matrix $\hat{\mathbf{X}}$, i.e., a PCA model, with dimensions IxJ. The structure of $\hat{\mathbf{X}}$ can be defined as:

$$\hat{\mathbf{X}} = \mathbf{A}\mathbf{B}^T \tag{4.1}$$

where $\mathbf{A}$ has dimension I x F and $\mathbf{B}$ has dimension J x F and $^{\mathbf{T}}$ indicates the transpose. The matrices $\mathbf{A}$ and $\mathbf{B}$ hold the parameters of the PCA model and are called scores and loadings, respectively. The number of columns F in $\mathbf{A}$ and $\mathbf{B}$ is usually much lower than the original number of variables in $\mathbf{X}$, suggesting that $\mathbf{A}$ and $\mathbf{B}$ give a compressed representation of $\mathbf{X}$. Each pair of f variables in $\mathbf{A}$ and $\mathbf{B}$ is called Principal Component, or component. The number of components to be included in a PCA model should be enough to extract all the relevant information filtering out uninteresting variation such as the noise. Since $\hat{\mathbf{X}}$ is a description of $\mathbf{X}$, the following relation can be formalized:

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} \tag{4.2}$$

where $\mathbf{E}$ is the residual matrix with size IxJ that holds the unexplained part of $\mathbf{X}$ in $\hat{\mathbf{X}}$.

A PCA model is bilinear and an alternative, yet equivalent, notation, useful to introduce this concept, is:

$$\hat{x}_{ij} = a_{i1}b_{j1} + \cdots + a_{iF}b_{jF};\ \ i = 1, \cdots, I;\ \ j = 1, \cdots, J; \tag{4.3}$$

which is equal to:

$$\hat{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{a}_f \mathbf{b}_f^T \tag{4.4}$$

where $\mathbf{a}_f$ and $\mathbf{b}_f$ are the fth column vectors of $\mathbf{A}$ and $\mathbf{B}$. As it can be seen from equation 4.3; $\hat{x}_{ij}$ is both linear in $a_{i1},\cdots,a_{iF}$ fixing $b_{j1},\cdots,b_{jF}$ and in $b_{j1},\cdots,b_{jF}$ fixing $a_{i1},\cdots,a_{iF}$ (Kruskal, 1983). At the same time it implies that the main assumption is that the experimental signal can be described by a linear model and this assumption is, at least theoretically, met for GC-MS data (Booksh and Kowalski, 1994). The concept of linearity can be extended to trilinearity and more, meaning that linear models can be defined also for data with more than two dimensions. All the models described in this section are bi or trilinear. From equation 4.4 a PCA model $\hat{\mathbf{X}}$ can be seen as the outer product of the matrices $\mathbf{A}$ and $\mathbf{B}$.

The columns of $\mathbf{A}$ and $\mathbf{B}$ are subject to the orthogonality constraint. Constraint(s) are imposed to a model to ensure specific properties or according to *a priori* knowledge. The orthogonality costraint determines that each component is 'blind' to the others, i.e., each component is explaining a specific part of the data and the same part is not explained by any other component. The orthogonality constraint is imposed on the loadings and the scores and it can be formalized as:

$$\mathbf{A^T A} = \mathbf{D};\ \mathbf{B^T B} = \mathbf{I} \tag{4.5}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{I}$ is an identity matrix due to the fact that the vectors in $\mathbf{B}$ are normalized.

The loss function is another fundamental aspect of a model; it defines how well the model approximates the data and also is used as cost function to estimate the parameters of the model. For a PCA model, the loss function can be formalized as:

$$\min ||\mathbf{E}||_F^2;\ \text{where } \mathbf{E} = \mathbf{X} - \mathbf{A}\mathbf{B}^T \tag{4.6}$$

where $||\ \ ||_F$ is the Frobenius norm of a matrix, i.e., the square root of the sum of the squares of the entries in equation 4.6. There are different algorithms to calculate the parameters in $\mathbf{A}$ and $\mathbf{B}$ for a PCA model, and nowadays the most applied is the singular value decomposition (Golub and Kahan, 1965). However, it is worth introducing another approach for calculating PCA, which is applied also to all the models described in this section. This approach is called Alternative Least Squares

(ALS) algorithm. A generic ALS algorithm estimates $\mathbf{A}$ and $\mathbf{B}$ to minimize the loss function in equation 4.6, with a multi step approach. The basic idea is to estimate, in a least square sense, $\mathbf{B}$ using a tentative initialization of $\mathbf{A}$. This can be formalized as follow:

$$\mathbf{B} = \mathbf{X}(\mathbf{A}^{+})^{T} \tag{4.7}$$

where $\mathbf{A}^{+}$ is the Moore-Penrose inverse of $\mathbf{A}$.

The estimation of $\mathbf{B}$ can be used to estimate $\mathbf{A}$ and so on, with an iterative approach until convergence. (Young, Takane, and Leeuw, 1978).

The basic principle of the unconstrained ALS can be outlined as follow:

1. initialize $\mathbf{A}_f$;
2. solve $\mathbf{B} = \mathbf{X}(\mathbf{A}^{+})^{T}$;
3. solve $\mathbf{A} = \mathbf{X}(\mathbf{B}^{+})^{T}$;
4. iterate through 2. and 3. until convergence;

The convergence of ALS has been demonstrated in De Leeuw, Young, and Takane, 1976. The convergence of the algorithm can be measured by comparing two consecutive iterations; when the difference in terms of fit is below a given threshold the convergence is reached.

One advantage of the ALS algorithm is the possibility to easily include constraints. For PCA models, an example of ALS based algorithm is the Non Iterative PArtial Least Squares (NIPALS) (Wold, 1975). This variation of the ALS alghorithm serves to impose the orthogonality constraint to the PCA model.

The orthogonality constraint is useful in a number of situations, but for the analysis of GC-MS data it can be problematic.

Considering the results from a GC-MS experiment for a single sample, the rows of the matrix hold the scans, i.e., the elution time, while the columns correspond to the measured mass spectra within the selected m/z ratios.

Keeping in mind the purpose of resolving pure signals of the different compounds in

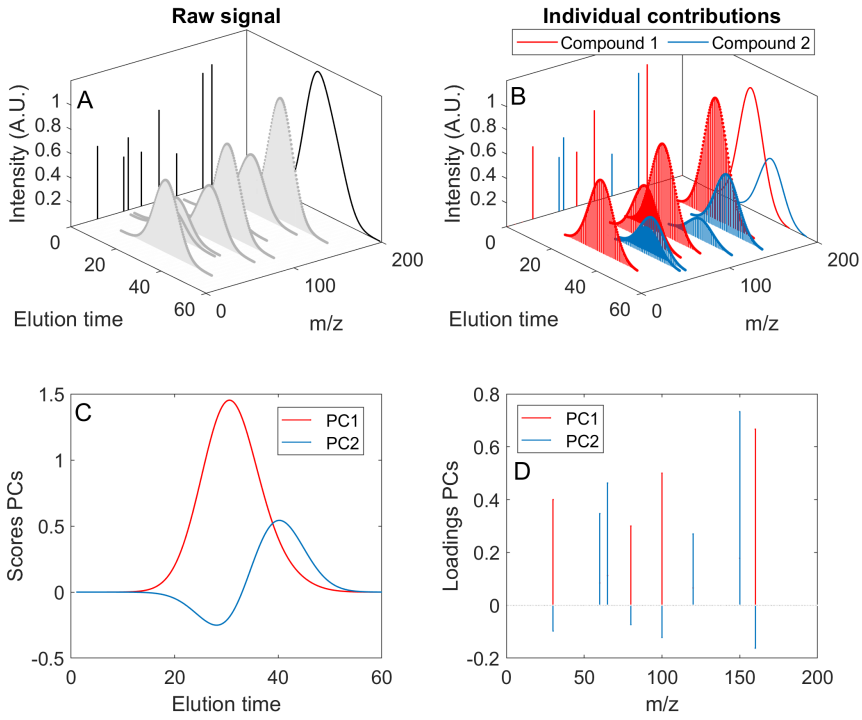

Figure 4.1 – A graphical example showing the solutions by a PCA model on GC-MS data.
A: Total signal from the GC-MS experiment. The TIC (black curves), the 3D raw data (grey surface), and the m/z (black lines) are shown
B: The same signal is decomposed in the individual contributions showing that two compounds are eluting at the same time.
C: Score plot of 2 components PCA model.
D: Loadings from the two components PCA model.

the raw data, the scores and loadings should be organized as follows. Each column vector of the scores of a PCA model would hold the total intensity of signal for each scan for a given compound. It can be seen as the specific TIC for the corresponding compound described by a given component. Each column vector from the loadings would hold the specific mass spectra for a specific compound. The scores and the loadings can be subsequently used to quantify and identify the resolved chemical signals, integrating the area under the peaks and comparing the resolved spectra with libraries, respectively.

Nonetheless, the orthogonality constraint imposes that each resolved signal is orthogonal to each other, meaning that overlapping signals cannot be properly resolved.

In figure 4.1 the loadings and the scores from a PCA two components model built on GC-MS data are shown. The number of components is consistent with the chemical rank, thus it should be sufficient to optimally describe the data represented in figure

4.1.A. It is possible to notice that the scores from PC1 describe the mean TIC shape, while the scores from PC2 are characterized by a first derivative shape, due to the orthogonality constraint, addressing for the shift due to the coelution of the two compounds. This is a typical shape for PCA components describing shifted data (Malmquist and Danielsson, 1994; Wülfert, Kok, and Smilde, 1998; Stoyanova and Brown, 2001), limiting the interpretability and chemical usefulness of PCA for the analysis of raw GC-MS data. Also looking the loadings, describing the mass spectra, the second PC fails in the extraction of chemical information, showing negative signals which are not present in the experimental spectra. In general, it is possible to obtain overlapping signals in untargeted metabolomics, thus imposing this constraint does not allow a proper representation of the data, making the PCA model not suitable for the analysis of GC-MS data.

4.2.2 MCR

Multivariate curve resolution (MCR) is another modeling approach. A MCR model decomposes a matrix $\mathbf{X}$ into a set of two smaller matrices $\mathbf{C}$ and $\mathbf{S}$ holding the concentration profile and the pure spectra, similarly to PCA. A MCR model can be formalized as follow:

$$\hat{\mathbf{X}} = \mathbf{C}\mathbf{S}^T. \tag{4.8}$$

where $\hat{\mathbf{X}}$, with dimension IxJ, is the MCR model, $\mathbf{C}$, with dimensions IxF, is the matrix holding the concentration profiles (scores) and $\mathbf{S}$, with dimensions JxF, is the matrix holding the pure spectra (loadings).

The main difference between PCA and MCR is that the orthogonality constraint is not imposed to the model. This is relevant because it implies that the signals described by the model can be overlapped, thus allowing a more pertinent representation of the original data.

In the context of the analysis of GC-MS data, the aim of applying a modelling approach is that each component describes a specific contribution, thus for each contribution a component should be included. This is called chemical rank (Smilde, Bro, and Geladi, 2005). For instance supposing that the data in a matrix holds the signals made of the contributions from two chemical compounds, the chemical rank of the

matrix is two, and a model including two components should extract all the relevant chemical variation, filtering out the irrelevant contributions such as the experimental noise. Thus, the number of components F included in the model should be related to the chemical rank of the original data. The same relation as in equation 4.2 can be assumed also for a MCR model. Hence, as for PCA, the number of F columns in $\mathbf{C}$ and $\mathbf{S}$ is much lower than J.

The ALS algorithm can be used to find the parameters of the model in $\mathbf{C}$ and $\mathbf{S}$ (**de2009two**). During the calculation of a MCR model each the f column in $\mathbf{C}$ and $\mathbf{S}$ are calculated at the same time, as such MCR is a non nested model. It means that the $F-1$ components model is not a subset of the F components model, thus the shape and the variance explained by the components can change when the number of components included changes. The algorithm can be formalized as follows (Tauler, 1995):

1. Initialize $\mathbf{C}$;
2. Solve $\mathbf{S} = \mathbf{X}(\mathbf{C}^{+})^{T}$;
3. Solve $\mathbf{C} = \mathbf{X}(\mathbf{S}^{+})^{T}$;
4. Iterate through 2. and 3. until convergence.

Different strategies for the initialization can be used, details on this aspect and in general about the algorithmic side can be found in (Tauler, 1995; De Juan and Tauler, 2006). Not imposing the orthogonality constraint, the components can effectively describe the specific contributions in a signal, as shown in figure 4.2, where a two components MCR model is represented. The number of components is consistent with the chemical rank of the data in 4.2.A, and the MCR model extracts the same profiles as in 4.2.B. Looking the scores in 4.2.C, Component 1 describes Compound 1, and Component 2 describes Compound 2. Finally, comparing the deconvolved spectra in 4.2.D with the real spectra in 4.2.B, no significant differences can be noticed.
Due to its effectiveness, MCR has been widely applied for the resolution of complex mixtures (De Juan, Jaumot, and Tauler, 2014).

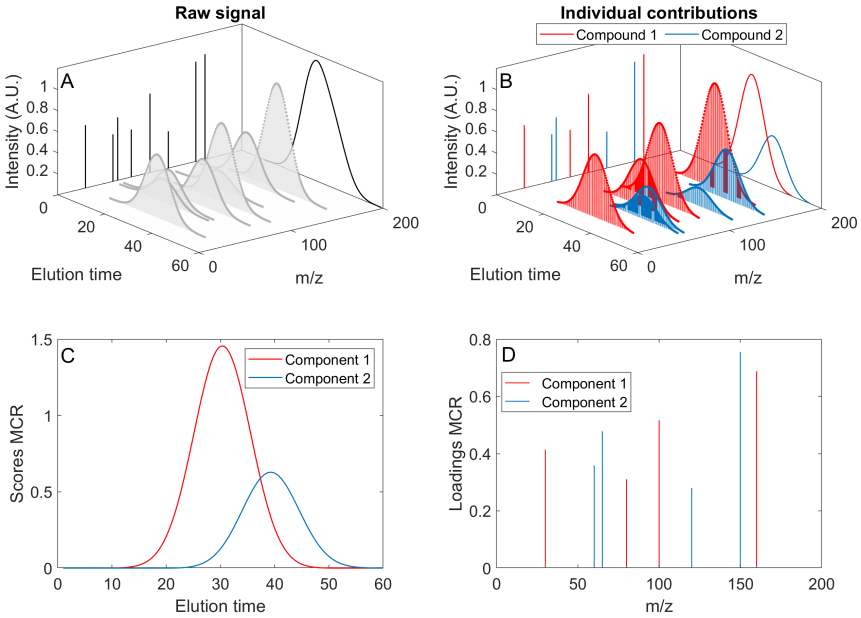

Figure 4.2 – A graphical example to show the solutions by a MCR model on GC-MS data (same as figure 4.1). A: Total signal from the GC-MS experiment. The TIC (black curves), the 3D raw data (grey surface), and the m/z (black lines) are shown
B: The same signal is decomposed in the individual contributions revealing two compounds are eluting at the same time.
C: Scores plot of 2 components MCR model. D: Loadings from the two components MCR model.

However, despite the effectiveness of the MCR model to resolve signals, there are some aspects that are worth noting.

The first one is that MCR can be applied only on two way datasets. When more than one sample is analyzed with GC-MS, the resulting data is actually three way. When a GC-MS experiment is performed, a matrix with dimensions IxJ is obtained for each sample. When K samples are analyzed, the resulting three way matrix $\underline{\mathbf{X}}$ has IxJxK structure. Thus, before to apply MCR the $\underline{\mathbf{X}}$ matrix must be unfolded (Garreta-Lara et al., 2016), leading to a huge increase of the matrix dimensions as shown in figure 4.3.

Another more important point to be considered is that, when applying MCR, the solution is not unique, due to the so called rotation ambiguity (Abdollahi and Tauler, 2011; Olivieri, 2021). Rotation ambiguity can be formalized as follows:

$$\hat{\mathbf{X}} = \mathbf{C}\mathbf{T}^{-1}\mathbf{T}\mathbf{S}^{\mathbf{T}} \tag{4.9}$$

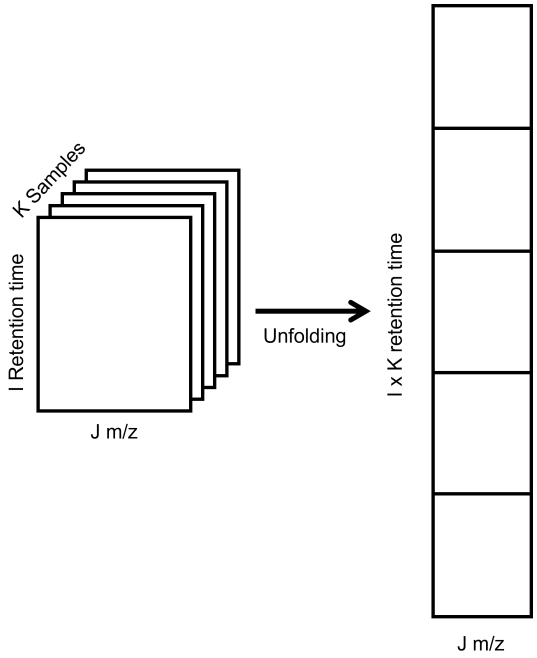

Figure 4.3 – Unfolding of a three-way matrix into a two-way matrix.

where $\mathbf{T}$ (FxF) is a non-singular matrix. Basically equation 4.9 shows that any rotation of the matrices $\mathbf{C}$ and $\mathbf{S}$ by means of $\mathbf{T}$ will produce the same solution, thus an infinite number of solutions with same fit can be obtained. It means that MCR can find the 'true' solution, i.e., the combination of parameters that actually describes the physical phenomena in the data. However, these parameters have to be find in a set of infinite combinations that describe the data equally well but not extracting the real contributions.

This issue, well known since when the method was proposed (Lawton and Sylvestre, 1971), can be tackled by imposing constraints to the model, thus reducing the number of feasible solutions (Abdollahi and Tauler, 2011). Another possible way is to take advantage of the three way data structure from GC-MS experiments.

4.2.3 PARAFAC

In 1944, Cattel proposed the principle of parallel proportional profiles (PPP) summarized as follows by himself in a later publication:

The basic assumption is that, if a factor corresponds to some real organic

> *unity, then from one study to another it will retain its pattern, simultaneously raising or lowering all its loadings according to the magnitude of the role of that factor under the different experimental conditions of the second study.* No inorganic factor a mere mathematical abstraction, would behave this way ... This principle suggests that every factor analytic investigation should be carried out on at least two samples, under conditions differing in the extent to which the same psychological factors .. might be expected to be involved. We could then anticipate finding "true" factors by locating the unique rotational position (simultaneously in both studies) in which each factor in the first study is found to have loadings which are proportional to (or some simple function of) those in the second: that is to say, a position should be discoverable in which the factor in the second study will have a pattern which is the same as the first, but stepped up or down. (Cattell and Cattell, 1955) Author italic"

A way to show the basic idea behind this principle is given in (Bro, 1998), and reported here.
A matrix $\mathbf{X}_1$ can be decomposed in two matrices $\mathbf{A}$ and $\mathbf{B}$. Supposing that $\mathbf{A}$ and $\mathbf{B}$ have two columns, the following relation can be formalized:

$$\mathbf{X}_1 = \mathbf{a}_1\mathbf{b}_1^T c_{11} + \mathbf{a}_2\mathbf{b}_2^T c_{12} \tag{4.10}$$

where c_{11} and c_{12} are equal to one. Another matrix $\mathbf{X}_2$ can be decomposed with the same parameters in $\mathbf{A}$ and $\mathbf{B}$ but with different proportions:

$$\mathbf{X}_2 = \mathbf{a}_1\mathbf{b}_1^T c_{21} + \mathbf{a}_2\mathbf{b}_2^T c_{22} \tag{4.11}$$

where c_{21} and c_{22} are not equal to c_{11} and c_{12}. This example summarises the idea behind the Parallel Proportional Profiling (PPP). Moreover, Cattel shown that the presence of parallel profiles would led to unambiguous decomposition. It implies that, if a model is based on PPP, it is not possible to rotate the model and obtain the same fit, resolving the rotation ambiguity that affects two way models, such as MCR.
Starting from Cattels' intuition, Harshman developed the PARAllel FACtor analysis

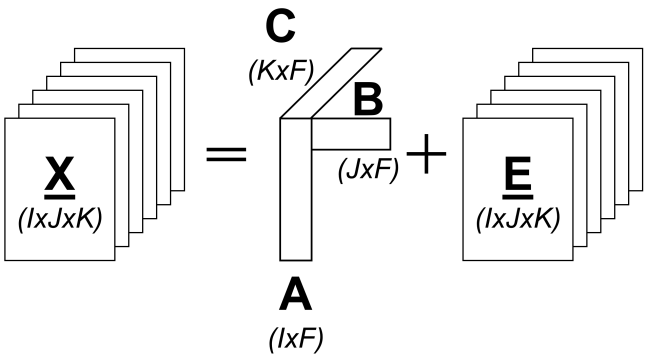

Figure 4.4 – Graphical representation of a PARAFAC model.

model (PARAFAC) (HARSHMAN, 1970).

The PARAFAC model is strictly related to the previous example and its structure is depicted in figure 4.4. Different equivalent notations can describe a PARAFAC model, the matrix sample-wise notation being among the most common ones:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T + \mathbf{E}_k \tag{4.12}$$

where $\mathbf{E}$ accounts for the unexplained variation in $\mathbf{X}_k$; that is the k-th slab in $\mathbf{X}$ and $\mathbf{D}_k$ is a diagonal matrix holding the kth row of $\mathbf{C}$ in its diagonal. The outer product notation for PARAFAC can be expressed as:

$$x_{ijk} = \sum_{f=1}^{F} a_{if}b_{jf}c_{kf} + e_{ijk}; \tag{4.13}$$

it can be seen from 4.13 that across the K samples the parameters in $\mathbf{A}$ and $\mathbf{B}$ are the same for all the samples, while the differences in terms of magnitude are given by the values in $\mathbf{D}_k$.

ALS can be used to calculate a PARAFAC model (Bro, 1998). The optimization problem is defined using the Kathri-Rao product notation and each loading matrix is estimated using the other two. PARAFAC is a non nested model and all the components are calculated at the same time during the fitting of the model. Details about the algorithm are given in Bro, 1998.

When a PARAFAC model is calculated on a three way matrix, where the elution time, the masses and the samples are in the first, second and third mode respectively,

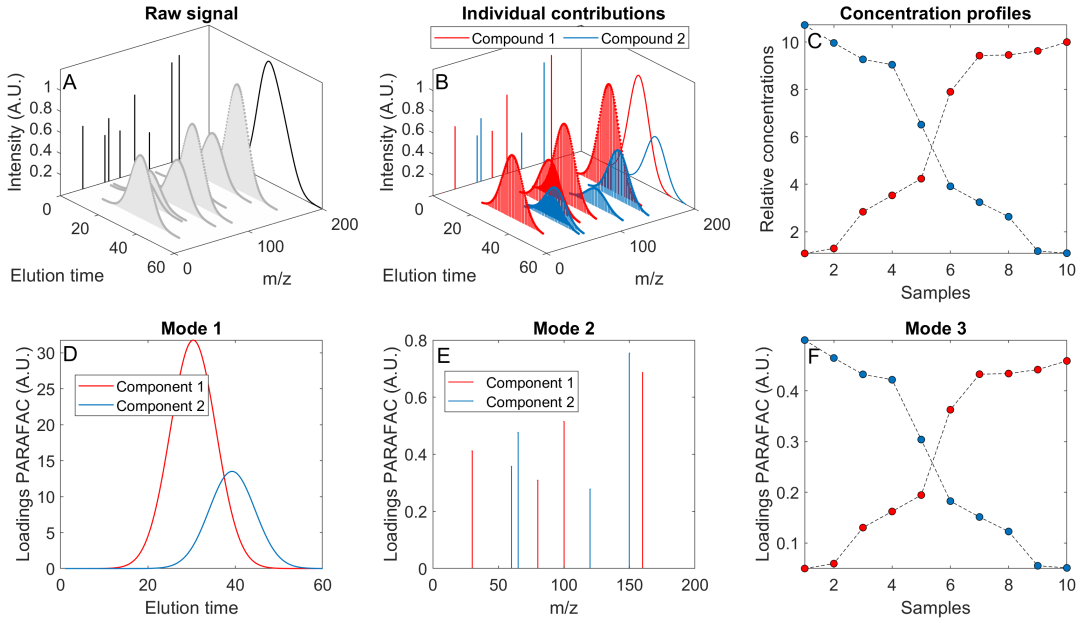

Figure 4.5 – A graphical example to show the solutions by a MCR model on GC-MS data (same as figure 4.1 and 4.2). A: Total signal from the GC-MS experiment. The TIC (black curves), the 3D raw data (grey surface), and the m/z (black lines) are shown
B: The same signal is decomposed in the individual contributions revealing two compounds are eluting at the same time.
C: Concentration profiles of 10 samples for the two compounds, in red the concentration profile for compound 1 in blue for compound 2.
D, E, F: Loadings from a 2 components PARAFAC model on the first second and third mode respectively. Notice the similarity of the loadings with the actual contributions in B and C

each column in **A**, **B** and **C** will describe the elution profiles, the mass spectra and the relative concentration of the samples, as shown in figure 4.5. By convention, the first two modes are called loadings and the third is called scores matrix. It means that, at least in principle, the individual contributions in a combination of signals can be resolved. Moreover, due to the uniqueness property, a PARAFAC model would decompose the data extracting the true contributions.

However, the data must fulfill some conditions in order to obtain a unique PARAFAC model. First, the dataset must contains at least two samples with independent variations in the concentration for a given compound. It means that just two replicates of the same sample can not be used, but at the same time it is worth to note that, when, among samples, some compounds have different concentrations while others do not, it is still possible to find the unique patterns for the freely varying compounds (Kiers and Smilde, 1995).

Another important point to ensure the uniqueness of a PARAFAC model is that the

right number of components is included in the model. Proofs and generalization of the uniqueness for a PARAFAC model in relation with the number of components have been discussed in different publications, nonetheless a feasible analytical method for the estimation of the right number of components is not available and a number of publications are devoted to the study of this property (Ten Berge and Sidiropoulos, 2002; Rajkó et al., 2017; Giordani, Rocci, and Bove, 2020).

In practice, the assessment of a PARAFAC model is performed by visually inspecting the results and using diagnostic tools on the calculated model. The two most common are the split half analysis (Harshman and Lundy, 1984) and the CORe CONsistency DIAgonstic (CORCONDIA) (Bro and Kiers, 2003).
Split half analysis consists in the repeated fitting of the same PARAFAC model considering only a fraction of samples. The main idea is to see if the same model, considering different subsets of samples, can be obtained. If this is the case, it can be considered as an evidence that the estimated components are reflecting true phenomena. Split half analysis is a powerful technique, but it is also time demanding.

CORCONDIA is very effective for the determination of cases where too many factors have been included in the model (Bro and Kiers, 2003). CORCONDIA estimates the appropriateness of a PARAFAC model by measuring if the variation explained by the model is low-rank trilinear, in other words if the components are explaining independent variation. For instance, if the samples consist of four chemical compounds and they are modelled with a five components PARAFAC model, the model is forced to divide the signals of the four compounds in five components, thus the components will not be independent to each other and the described variation is not low-rank trilinear. In these cases CORCONDIA will be low.

Another fundamental assumption for PARAFAC can be observed directly from equation 4.12. It is possible to notice that the F columns in $\mathbf{A}$ and $\mathbf{B}$ are equal for all the K samples. This means that the PARAFAC model assumes that the contribution of a specific factor is equal across all the samples. Theoretically, GC-MS data can be seen as a trilinear structure, thus satisfying the PARAFAC assumption. Nonetheless,

this is not always true with experimental data, due to two main problems: the shift in retention time for a given compound between samples, as well as changes in peak shape, again considering different samples. The deviations from trilinearity in GC-MS data mainly arise from the elution profiles, while the mass spectra obtained with EI in low resolution are more stable across the samples. However, problems can arise when the mass spectrometer detector is saturated and deviations from linearity can be observed also in the mass mode (Ausloos et al., 1999; DeJong et al., 2016).
These observations highlight the limits of PARAFAC to model GC-MS data. To overcome these limitations, data can be either aligned or unfolded and analyzed by means of MCR, loosing the uniqueness property of PARAFAC. Another option is to apply another modeling approach, able to handle the time shifts and the shape changes.

4.2.4 PARAFAC2

PARAFAC2, proposed by Harshman (Harshman et al., 1972), is related to the PARAFAC model and equation 4.12 can be used to introduce the model. PARAFAC2 can deal with shifting modes replacing the $\mathbf{B}$ matrix with a set of $\mathbf{B}_k$ matrices. Thus the model can be rewritten as:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^T + \mathbf{E}_k \tag{4.14}$$

where $\mathbf{B}_k$ has dimension $J\text{x}F$. Intuitively, as it is written in equation 4.14, the model is not consistent with the the principle of parallel proportional profiles, since it requires that the profiles are consistent between different samples. If this condition is not satisfied, then the model would lose the uniqueness property. To maintain this property, Harsman imposed the following constraint:

$$\mathbf{B}_k^T\mathbf{B}_k = \mathbf{H}, \quad \text{with } k = 1, \ldots, \mathbf{K} \tag{4.15}$$

where $\mathbf{H}$, is a $F\text{x}F$ matrix invariant across the k slabs. For 4.15 to hold each $\mathbf{B}_k$ is decomposed as:

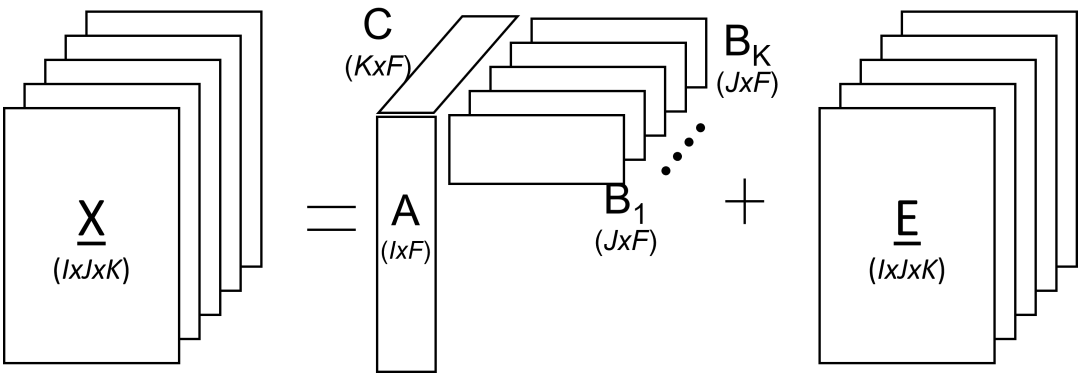

Figure 4.6 – Graphical representation of a PARAFAC2 model.

$$\begin{aligned} \mathbf{B}_k &= \mathbf{P}_k\mathbf{H}, \quad \text{for } k = 1, \ldots, \mathbf{K} \\ \mathbf{P}_k^T\mathbf{P}_k &= \mathbf{I} \end{aligned} \tag{4.16}$$

where $\mathbf{P}_k$ is a JxF matrix and $\mathbf{I}$ is a FxF identity matrix, suggesting that $\mathbf{P}_k$ is an orthonormal matrix, i.e., normalized orthogonal matrix. With these constraints the model is unique under mild conditions (Berge and Kiers, 1996). A graphical interpretation of the model and an example to show the differences between PARAFAC and PARAFAC2 are given in figures 4.6 and 4.7, respectively. In 4.7.B, 4.7.C and 4.7.D it possible to notice that the components from a PARAFAC model cannot effectively describe the data shown in 4.7.A, due to the shifts and shape changes across the samples. Instead, a PARAFAC2 model can handle the shift, resulting in a clear description of the data, as shown in figure 4.7.A. In figure 4.7.E resulting TIC from the components obtained with PARAFAC2 are represented and it is possible to notice that the specific contributions have been effectively separated.

As for PARAFAC, the uniqueness is obtained when the right number of components is included in a PARAFAC2 model. Also in this case the concept of chemical rank is important. A visual inspection of the model was one of the main way to asses the right number of components together with the analysis of fit of the model, the number of iterations etc.. However, more sophisticated tools have been developed to assist the selection of the models with the right number of components. For instance, CORCONDIA has been adapted for PARAFAC2 (Kamstrup-Nielsen, Johnsen, and Bro, 2013), and also for this model proved to be a valuable tool.

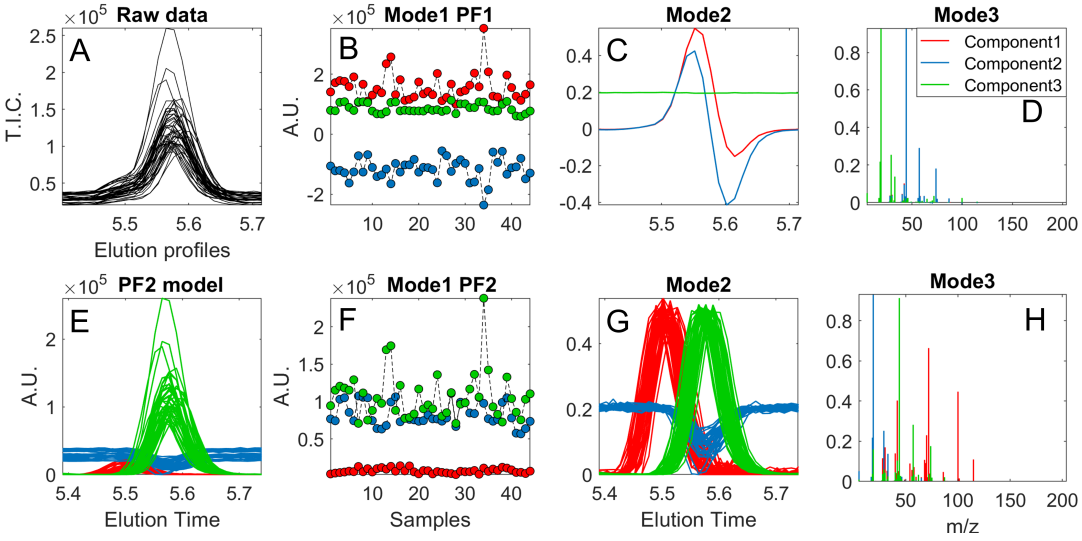

Figure 4.7 – A) Raw data extracted from the dataset in Ballabio et al., 2008. The visual inspection of the TIC suggests that a 3 components model would separate all the relevant chemical information: two coeluting compounds and the baseline.
B), C), D) first second and third mode loadings from a PARAFAC (PF1) model.
F), G), H) first second and third mode loadings from a PARAFAC2 (PF2) model.

Another tool is a Partial Least Squares - Discriminant Analysis model developed to identify models with the right number of components (Johnsen et al., 2014). The PLS-DA model was trained on diagnostic values estimated directly from the PARAFAC2 models, showing good performances.

Also for PARAFAC2 it is possible to impose constraints on the loadings and recently a modification of the original ALS-PARAFAC2 has been proposed to impose non negativity on all the three modes (Cohen and Bro, 2018). This flexible version of PARAFAC2 is important for two reasons:

1. the experimental data are always non negative, thus the models will fit the data in a more appropriate way;

2. it has been shown that the constraint reduces the number of local minima, thus reducing the uncertainty of the estimated models (Cohen and Bro, 2018; Yu and Bro, 2021).

4.2.5 PARAFAC2 for GC-MS data: approach and drawbacks

Given that a PARAFAC2 model can handle temporal shifts, it is a preferable modelling approach compared to PARAFAC, since no or minor preprocessing is required for

modeling GC-MS data (Amigo, Skov, and Bro, 2010). PARAFAC2 has been successfully applied for the analysis of GC-MS data; Yuille et al., 2018; Sales et al., 2019; Laursen et al., 2020 are an example of some of the most recent applications. The release of the PARADISe platform (Johnsen et al., 2017) has been a major breakthrough for the use of PARAFAC2. This is a GUI that removed the need of line coding for PARAFAC2 modeling. Given the uniqueness of the solutions, the possibility to handle chromatographic issues such as shifted peaks and the number of pubblications where it has been successfully applied, PARAFAC2 has been identified as the best available option for AutoDise system to resolve the specific contributions in the experimental GC-MS data.

In practice, some steps are required to apply PARAFAC2 and PARADISe is the first platform with a dedicated GUI to easily handle them. The steps are:

Interval definition PARAFAC2 models are calculated on specific regions defined on the elution time dimension of GC-MS data, called intervals. Usually the definition of the intervals is performed manually. Practically, if the aim is to resolve all the compounds, each peak in the data must be included in at least one interval. Usually a good practice is to define small intervals including a limited number of peaks, ideally one at a time. A properly defined interval should include a whole peak, crucial for an accurate quantification of the compound. With complex datasets, where the number of compounds is up to thousands, this step can be time consuming. Moreover, the peaks hidden in the baseline could be undetected by the user. Furthermore, in real metabolomics GC-MS experiments, coelution of compounds is extremely frequent and often it is not clear where to put the limits of the intervals. This can be solved later by analyzing the models and defining new intervals, but a thorough and time consuming analysis of the models is required.

Models calculation Once the intervals are defined, the PARAFAC2 models are calculated. As described before, the user has to set the number of components included in each model. However, tools for a reliable assessment of the number of components to be included in the models before the calculation have not been yet proposed. The common approach is to calculate a set of models with

different number of components, and then the best solution is selected by visually inspecting the obtained results. The range of components included depends on the intervals. If the criteria to include a small number of peaks in each interval is adopted, a small chemical rank can be assumed and usually no more than 10 components are needed to obtain appropriate solutions.

Components selection The final step is the selection of the components to be included in the peak table, where for each sample the relative concentration of the resolved compounds is reported together with a tentative identification, obtained with the comparison of the resolved mass spectra with spectral libraries such as NIST14.

This is performed by visually inspecting the models and extracting the interesting components, i.e., the components that are describing the peaks. To assist the researchers during this phase, in Johnsen et al., 2014 and Risum and Bro, 2019 two effective tools have been proposed and described. In the first paper a PLS-DA model was developed to identify the appropriate models in terms of number of components.

In the second, a Convolutional Neural Network has been proposed to identify the components that are describing the peak. The network is able to label the elution profiles according to their shape. When the network is applied on the signals resolved by PARAFAC2, it can reduce the number of models to be analyzed by selecting only those where the peaks are described.

Nevertheless, these tools do not exclude the visual analysis of models, although their use is of great help during this step. The main point during this step is to extract and include in the peak table each of the resolved peaks only a single time, otherwise redundant compounds will be present. When PARAFAC2 is applied on complex GC-MS data, it is common to have overlapping intervals, thus models from different intervals can describe the same compound. The tools described above, when applied on PARAFAC2 solutions, do not select the components to be included in the peak table. The PLS-DA model give a list of models, and for each model the components describing the peaks have to be manually selected. The convolutional neural network provides a list of

components that are describing the peak. However, in both the cases, it is likely that some models/components are describing the same chemical compounds, thus the final selection is left to the user. The selection of the components is time consuming and can lead to biased and redundant peak table. Once the components have been selected, the relative quantification can be performed directly by the model, while the identification is performed by comparing the resolved spectra of the selected components with spectral libraries.

AutoDise, the algorithm developed during this PhD, described in chapter 5 of this thesis, aims to automatically handle these steps, reducing the analysis time and producing a comprehensive and non redundant peak table, with no actions required to the user.

4.3 Peak alignment

PARAFAC2 can handle time shifts in the data as well as baseline drifts. Therefore, GC-MS data do not need to be preprocessed before calculating a model. Nonetheless, for practical reasons, it can be useful to align the data, considering the elution time dimension. This is important during the definition of the intervals. If several samples are analyzed and the resulting data are affected by severe time shifts the definition of the limits of the intervals can be extremely complicated. For instance, peaks describing different molecules could be included in the same interval, while peaks referring to the same compound can be split in different intervals. These issues can be detected afterward during the visual inspection of the models. Nonetheless, a even non perfect alignment of the data can be extremely useful in terms of time saving both in the interval definition step and in the visual inspection of the models. Here two effective alignment approaches are briefly described: CoShift and COW.

4.3.1 CoShift

CoShift has been proposed by van den Berg and colleagues in (Engelsen, Belton, and Jakobsen, 2005). This is a rigid alignment algorithm, the idea is to horizontally shift a chromatogram to maximise the correlation with a reference profile.

The reference is of course important to obtain good performances, and it must be as

generic as possible so to have with all the samples some common peaks. In Skov et al., 2006, four methods have been suggested to select the reference, but the final choice is left to the user. The types of references are:

- the mean of the signals;
- the median of the signals;
- the bi-weighted mean of the signals;
- the maximum of the signals;
- the maximum cumulative product of correlation coefficients.

The mean, the median and the bi-weighted mean are similar to each other, overall the references produced by these approaches are characterised by wider peaks than in the raw data, particularly for the mean reference, while using the median usually the peaks are narrower. Using the bi-weighted mean, the outliers are down weighted in the calculation of the mean, the resulting reference is an average between the two others. The maximum signal is defined by the maximum values at each elution time across all the samples, this option is rarely applied. The maximum cumulative product of correlation coefficients instead selects a real chromatogram among those in data, in particular the profile from the sample with the highest correlation with all the others is selected.
Intuitively, this last option is particularly suited for this approach. The step size to find the maximum correlation is limited by a defined threshold, to avoid the alignment of peaks unrelated to each other.

4.3.2 COW

COW stands for Correlated Optimized Warping (Tomasi, Van Den Berg, and Andersson, 2004). The main idea is to align all chromatograms in a batch towards a reference chromatogram, one chromatogram at a time. The reference is of crucial importance to obtain an effective alignment. The same options mean, median, bi-weighted mean, maximum intensity and the maximum correlation, described for CoShift can be also applied to define a reference for COW.

Once the reference is defined, the reference and the chromatogram profiles are divided into intervals which are to be aligned. The intervals can vary, so they can be enlarged or shrunk according to a threshold of flexibility defined by the slack size parameter, which indicates the maximum length change. Afterwards, all possible combinations of segment sizes are tested. The combination of segment size, which results in the best alignment accordingly to the defined slack sizes, is chosen. If the segment of the chromatogram to be aligned has a different size compared to the corresponding segment in the reference, it is linearly interpolated to fit the size in the reference.

The length of the intervals and the slack size are crucial to obtain the best alignment, ensuring at the same time enough flexibility to correct the shifts, but avoiding any changes in peak shape'. Skov and colleagues (Skov et al., 2006) have proposed an automated way to optimize these parameters.
The optimisation tests combinations of different slack and segment sizes, within a defined range, with a simplex based approach to find the optimal parameters. The optimal parameters are defined by means of two criteria, used to assess the goodness of the alignment: simplicity and peak factor.
The simplicity value is the fourth power of the first singular values of the aligned chromatogram. The more the simplicity is close to one the less the different profiles are shifted, thus indicating that the profiles have been aligned.
The peak factor is based on the difference in the Euclidean norm of the sample chromatograms before and after the alignment. If the norm of the chromatographic profile changes after the alignment, as a result of the interpolation, it will influence the area under the peak. Thus the peak factor is important to verify if the peak shapes are preserved. Through the automatic optimisation, the amount of manual work to align chromatograms is extremely limited, nonetheless the optimisation step requires a significant amount of time.

CoShift is more suited to align chromatographic profiles with major shifts, instead when the shifts are limited and there is no need to introduce too much flexibility, which can dramatically affect the shape of the peaks, COW is the preferable option. Thus,

the two approaches can be combined: in a first round CoShift is applied to correct the biggest shifts, then COW solves the remaining (minor) issues. Solving the major shifts with CoShift is an advantage because COW would need too much flexibility, if the shifts are severe, destroying the shape of the peaks.

4.4 Artificial neural networks

Artificial neural networks (ANN) are a machine learning approach roughly inspired by how neurons in a human brain work. In the recent years, the application of neural networks has increased thanks to the huge availability of data and fast improving in computing capacity.

In this thesis, two types of ANN have been tested: Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). Shortly some concepts will be introduced by describing the basics of ANN and then details will be given on RNN and CNN. These methods are a crucial elements for the AutoDise system. In particular, they have been applied to identify the PARAFAC2 components that describe the peaks in the data. As described in chapter 3 common approaches for the identification of the peaks are based on Gaussian filters applied on the chromatographic profiles. However, most of the times chromatographic peaks are not characterized by Gaussian shapes. Moreover, these approaches require to set parameters such as the width of the filter, which can dramatically affect the output. ANN approaches have been successfully applied in Risum and Bro, 2019 in this context, and the trained model has been implemented in AutoDise. However, during this project other networks have been trained, in order to test different strategies and eventually improve the labelling results. Details about the trained models are given in chapter 5, here a brief overview of the theory at the basis of these approaches is given.

4.4.1 Artificial Neural Network

An ANN can be defined as a function optimized to map an input $\mathbf{x}$ to a label y (Goodfellow, Bengio, and Courville, 2016). The function is usually described using

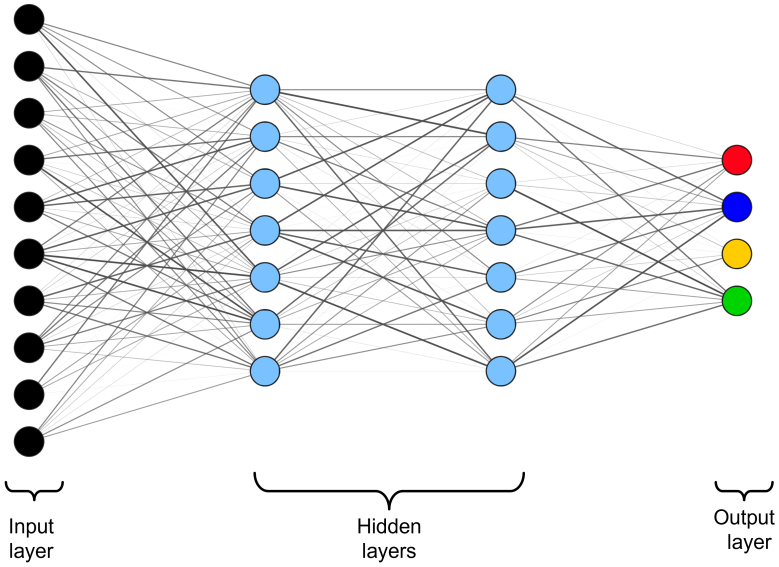

Figure 4.8 – Graphical representation of a ANN. The edges width and opacity is proportional to the respective weights. The different colours of the output layer represent four different classes.

neurons and weights.

In each neuron of an ANN there are inputs and one output. The relation between the inputs and the output can be formalized as follow:

$$y = f(w_1 x_1 + \cdots + w_j x_j + \cdots + w_J x_J + n) \tag{4.17}$$

where y is the output, w is the weight associated to the jth input x_j, $f()$ is the activation function and n is called bias, which is a threshold that influences the final output. In general, the greater is the bias of a given neuron the most likely the output of a neuron will be high in value. Different types of activation functions can be used, among the most applied are the sigmoid function, the ReLu function, and the tanh function (Sharma, Sharma, and Athaiya, 2017).

The neurons are organized in layers, the first layer is called input layer, the last layer is called output layer and all the layers in between are called hidden layers. If there is more than one hidden layer, then the network is called Deep Neural Network. In figure 4.8 a graphical representation of an ANN is shown. In the input layer, the sample features are fed into the network. The number of neurons in this layer depends on the number of features of the samples. If the samples have J variables, the number of

neurons in this layer will be J.

All the neurons in the input layer are connected with all the neurons in the first hidden layer. In the same way, all the neurons in the first layer are connected with all the neurons in the second layer, and so on. Connected means that the output of the previous neuron is transferred to the next one, so it is the input of the neurons of the subsequent layer. Then, all the information fed in each neuron is transformed as in equation 4.17 . Considering the layer structure of an ANN, 4.17 can be rearranged layer wise:

$$\mathbf{y}^{(l)} = f(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{n}^{(l)}) \tag{4.18}$$

where l is the layer number, $\mathbf{y}$ is the vector of the outputs for the lth layer with dimension (1xn), where n is the number of neurons in the lth layer, f is the activation function, $\mathbf{W}$ is the matrix of the weights with dimensions (kxn), where k is the number of inputs, $\mathbf{x}^{(l-1)}$ is the vector holding the outputs from the previous layer and $\mathbf{n}$ is the vector holding the biases of the neurons in the lth layer.

The last layer of neurons, called output layer, comprises as many neurons as the number of classes. Each neuron in this layer produces an output as the neurons in the hidden layers. A softmax operator is applied to the outputs. This is a function providing a normalized probability distribution over the possible classes defined as:

$$\hat{\mathbf{y}} = \frac{e^{\mathbf{x}^{(l-1)}}}{\sum_{k=1}^{K} e^{\mathbf{x}_k^{(l-1)}}} \tag{4.19}$$

where $\hat{\mathbf{y}}$ is the output of the softmax operator, $\mathbf{x}$ is the output from the previous layer and K is the number of classes. $\hat{\mathbf{y}}$ can be considered as the normalized probability distribution consisting of K probabilities for each sample.

The goal is to optimize the biases and the weights in the network so to minimize the cross entropy between the true distribution given in the training data and the predicted one. During the training, the data are fed into the network and the optimal biases and weights are estimated. This is performed by means of gradient descent algorithms, such as Stochastics Gradient Descent (SGD), Root Mean Squared Propagation (RMSProp) and Adaptive Movement Estimation (Adam). The parameters are optimized calculating

the gradient of the loss function by using the back-propagation algorithms; details about the algorithm can be found in Ruder, 2016.

4.4.2 Recurrent neural network

Recurrent Neural Networks (RNN) are a type of ANN specifically designed for sequential data, i.e., time series organized in a vector, with length I, where each element is a part of the sequence (Goodfellow, Bengio, and Courville, 2016). The core of RNN neural network is called 'cell'. A cell can be seen as a neuron with an internal loop. The internal loop allows the cell to keep the information from the previous input with the next input, which is the following element in the input vector. Thus, in a RNN each cell is fed with the whole sequence differently from ANN. The information from the previous time step is stored in the so called hidden state, which is the key to successfully classify sequential information. The hidden state can be formalized as:

$$h_t = tanh(\mathbf{W}h_{t-1} + \mathbf{W}x_t) \tag{4.20}$$

where h_t is the hidden state for x at the time t, $\mathbf{W}$ is the matrix of the weights, and h_{t-1} is the hidden state from the previous iteration considering x_{t-1}. From equation 4.20 it is possible to notice that the weights in $\mathbf{W}$ are the same for all the inputs, thus during the training the weights will be optimized considering the whole sequence. The output from the last iteration can be used to get the final classification.
The problem with shallow RNN is that the training, which is based on the calculation of the gradient of the loss function, can suffer from the so called vanishing gradient problem (Hochreiter, 1998). As described above, in each cell of a RNN there are as many layers as the length of the sequence. Because of the transformations of the *tanh* activation functions, during the training the gradient of the loss function, calculated by means of back propagation algorithm, can tend to zero, blocking the training of the model .

Long Short Term Memory cells have been introduced in 1997 (Hochreiter and Schmidhuber, 1997) to overcome this problem. There are four gates in LSTM cells, usually called i, f, o, g; these gates regulate the flow of the information within a cell

and are calculated with four activation functions, the first three gates by sigmoids functions and the latter by a tanh function.

These gates are used to calculate two hidden states, one defined as a hidden state and the other as cell state. Overall the process within an LSTM can be formalized as

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \mathbf{W} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f\dot{c}_{t-1} + i\dot{g} \tag{4.21}$$

$$h_t = o\dot{t}anh(c_t)$$

where i, f, o, g are the four gates, h_{t-1} is the hidden state for x the time $t-1$, $\mathbf{W}$ is the matrix of the weights, and x_t is the input at time t, c_t is the cell state at time t and h_t is the hidden state for the time t. The sizes are: ($4i_h$ x $2i_h$) for $\mathbf{W}$, ($4i_h$ x 1) for the vector holding the values of the four gates i, f, o and g and i_h correspond to the number of elements in the vector h_t.

The inner structure of a LSTM cell, in particular the forget gate, avoids the vanishing gradient problem during the back-propagation algorithm for the optimization of the weights in the network.

A modification of LSTM networks is called BIdirectional LSTM (BILSTM). A BILSTM network consists of two LSTM networks, one takes the input in the forward direction, the other in a backward direction, increasing the amount of available information to the network, i.e., the network will have information about both what precedes and follows a given input of the analyzed sequence.

4.4.3 Convolutional neural networks

Convolutional Neural Networks (CNN) are another type of ANN, widely applied for image classification (Goodfellow, Bengio, and Courville, 2016). The extraction of the patterns in the images is manly due to the use of filters or kernels.

A kernel can be seen as a shifting window moving over the image, which is the sample. The kernel is a matrix of weights. At each shift the weights are multiplied

element wise with the portion of the sample selected by the kernel. The result of this operation is called convolved matrix, which is a reduced form of the original matrix. The kernel and the calculation of the convolved matrix are done in the convolutional network, the core of CNN. The next layer is called pooling layer, which is another reduction performed on the convolved matrix, useful both in terms of decrease of computational cost and feature extraction.
A pooling layer is another kernel shifting over the convolved matrix. The two pooling strategies are: max pooling or average pooling. In the first case, the kernel will keep the maximum value of the windows over which the kernel is hovering, in the other case the values are averaged.

In CNN multiple layers can be stacked. For instance, AlexNet (Krizhevsky, Sutskever, and Hinton, 2012), one of the first GPU implemented CNN, winner of the ImageNet Large Scale Visual Recognition Challenge in 2012, has 5 convolutional layers and 3 pooling layers.
In classification contexts, the output from the convolutional layers is fed into a feed forward network to obtain the final classification. Also for CNN, the training goal is to optimize the kernel weights to minimize the loss function. The same optimizer and back propagation algorithm are also applied for CNN.

4.5 k-Nearest neighbours

k-Nearest neighbours (kNN) is a non-parametric modeling method used for both classification and regression purposes. During this project it has been applied for classification.
The basic idea is that the training samples are distributed in a space defined by their features. When a new sample has to be predicted by the algorithm, kNN uses the frequency of the classes of the k nearest neighbours included in the training set to determine the class of the new sample.
The comparison of new samples with those included in the training is made by calculating all the distances between the new (unknown) sample with all the others. Thus, one of the drawbacks of the approach is that the computation time needed to

classify a sample depends on the number of samples included in the training set, and it can be long with large databases. The major advantages of the KNN classification are its simplicity and the classification efficiency (Kowalski and Bender, 1972). The determination of the number of neighbors to use (k) is done by cross-validation.

4.6 Classification measures

The classification performances discussed in chapter 5 have been evaluated by means of confusion matrix and derived measures: Non-Error Rate (NER), precision, sensitivity and Receiver Operating Characteristic (ROC) curves.

The classification results can be summarized in the so-called confusion matrix. This is a $G \times G$ matrix, where G corresponds to the number of classes. Rows represent the experimental classes, while columns represent predictions. Thus, diagonal elements c_{gg} of the confusion matrix represent the number of correctly classified samples in the g-th class, while the off-diagonal values correspond to misclassification, such that c_{gc} holds the number of samples of class g predicted as class c. Given the confusion matrix, the sensitivity (Sn_g) for the g-th class is defined as:

$$Sn_g = \frac{c_{gg}}{n_g} \tag{4.22}$$

where c_{gg} is the number of samples of the g-th class correctly classified and n_g corresponds to the total number of samples that belong to the g-th class. As such, the sensitivity summarizes the ability of the model to identify the samples belonging to a given class. The NER, also known as Balanced Accuracy) is defined as the mean of class sensitivities

$$NER = \frac{\sum_{g=1}^{G} Sn_g}{G} \tag{4.23}$$

The precision (Pr_g) corresponds to the ratio of samples of class g correctly classified over the number of the samples predicted into the g-th class:

$$Pr_g = \frac{c_{gg}}{n\prime_g} \tag{4.24}$$

As such the precision is used to quantify how many of the samples predicted as

class g are actually belonging to that class.

Receiver Operating Characteristics (ROC) curves are a graphical tool for the diagnosis of a classification model. The curve for a given class g is obtained by plotting the False Positive Rate (FPR) versus Sn_g, also known as True Positive Rate (TPR), at various scores thresholds. The scores are calculated during the classification and indicate which is the most likely class for a given sample. Thus, it is possible to adjust the score thresholds for the assignment of the predicted class and calculate the respective FPR and TPR. The best model would be characterized by a point in the upper left corner of the plot, where specificity and sensitivity are both 100%, which correspond to FPR equal to 0 and TPR equal to 1. The Area Under the Curve (AUC) corresponds to the value of the area under the ROC curves. For instance, a perfect classification model has an AUC equal to 1 while for a random classifier, where the ROC curves correspond to the diagonal of the plot, the AUC would be 0.5. Further details for all the classification diagnostics can be found here (Ballabio, Grisoni, and Todeschini, 2018).

Chapter 5

Results

In this chapter the results of this PhD project are reported. The first part describes how the algorithm behind the AutoDise system has been designed and how it handles the steps to obtain a non redundant, yet comprehensive, peak table applying PARAFAC2 modeling on GC-MS data. Then the dedicated GUI, developed to implement AutoDise, is described. In the second part, the results from the deep learning and machine learning models for the classification of the elution profiles are reported.

5.1 AutoDise

AutoDise is an expert system, developed during this PhD project, to automatically obtain the peak table from the analysis of GC-MS untargeted data by means of PARAFAC2. It is based on a multistep workflow:

- intervals definition;
- models calculation;
- components selection;
- compounds identification and quantification;
- peak table.

5.1.1 Intervals definition

As described in chapter 4, PARAFAC2 models are calculated on defined elution time regions of the GC-MS data. This implies that if a peak is not (fully) included in any interval, it will not be resolved. Thus, in order to extract all the chemical information in the data, the aim is to define at least one interval for each peak. Many methods exist to identify the peaks in chromatographic data. For instance, in all the software described in chapter 3, peak detection algorithms are implemented. However, they are not automatic since the user is required to set different parameters, for example the width of the filter applied for the identification of the peaks. Another option to automate the definition of the intervals could be the AI binning proposed by De Meyer and colleagues (De Meyer et al., 2008), which detects the regions where there are peaks, would seem a feasible method. However, to apply these binning approaches and other related methods (Anderson et al., 2008; Davis et al., 2007; Anderson et al., 2011):

- the elution profiles have to be aligned;
- the boundaries are set at the local minima, i.e. the saddle points, between peaks.

One of the advantages of PARAFAC2 is the possibility to handle one shifting mode. Accurate profile alignment would make it unneeded to apply PARAFAC2.
Setting the boundaries at the saddle points is similar to the drop valley method. As described in chapter 2, it would lead to inaccurate quantification of the compounds. In a few words, setting the boundaries of the intervals at the saddle points, the tails of the peaks would be excluded, thus the integration of the area would provide misleading results. Moreover, since the boundaries are set at the saddle points, the peak must be detectable inspecting the raw data. It can happen that small peaks are hidden in the signal of the baseline, thus, applying these binning approaches, such peaks would not be detected and not included in any interval. These aspects limit the application of current approaches for the global analysis of raw GC-MS data.
Summing up, the definition of the intervals must take into account:

- the time shifts;
- the presence of unnoticeable peaks.

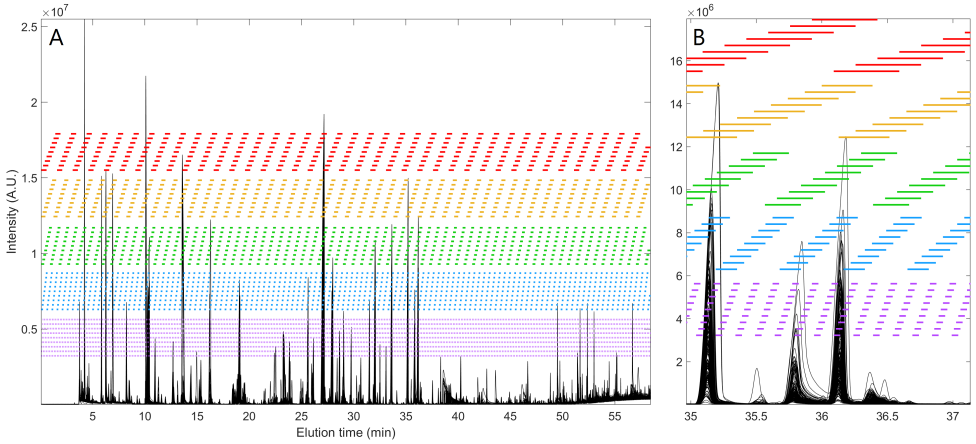

Figure 5.1 – A: In black are shown the TICs from Quintanilla-Casas et al., 2020, the coloured lines represent the intervals defined by AutoDise. Intervals with the same length are represented with the same color. The y position of the lines has the purpose to ease the visualization. B: Detail of the TIC and the intervals.

The interval definition implemented in AutoDise starts with a preliminary alignment which solves the most severe shifts. The aim of this step is to more easily assign which peaks belong together in the next steps. This step is only essential e.g. for runs that extend over months where the retention time shifts can be quite severe and abrupt and it is emphasized that the alignment is crude and does not aim at perfect alignment – just adjustment of severe shifts. Afterward, the intervals are set to span the whole length of the chromatographic run, i.e. the entire chromatogram is considered, with no region excluded. In particular, intervals of five evenly distributed lengths (in seconds) are defined. The minimum length is three seconds, the maximum is thirty, which is a common time range for the elution of peaks in GC-MS. For each of these interval lengths, the whole chromatogram is divided into intervals with the corresponding length from the first time point and onwards, such that these intervals overlap by two thirds. The result is a set of overlapping intervals spanning the whole chromatogram. As a result, about 85 intervals are defined for each minute of the chromatogram. As is possible to notice from figure 5.1, each part of the elution profiles is covered by several intervals. This over-redundant approach to select intervals has demonstrated to be important for two reasons:

- Extracts all the peaks including those not visually observable
- Obtains a proper interval for each peak in the data.

Intuitively, the computational time to fit the PARAFAC2 models will increase by defining so many intervals. Similarly, it can be expected to obtain several models that encompass the same chemical compounds. This will be exploited in the next steps as described below.

5.1.2 Models calculation

The model fitting is performed by means of the flexible PARAFAC2 algorithm described in Cohen and Bro, 2018. Differently fromthe normal PARAFAC2 algorithm based on ALS (Kiers and Smilde, 1995), the flexible approach allows non negativity constraint on each mode, including the resolved elution profiles. This is relevant for different reasons:

- the experimental data is non negative, thus the non negativity constraint gives a more accurate representation;
- numerical problems such as two component degeneracy are solved to a large extent;
- mostly ALS and flexible solutions are similar but when they are not, the flexible solution is mostly the better one, as shown in figure 5.2.

For each interval several models are calculated, with an increasing number of components. By default, the minimum number of components is one, the maximum is eight.

The number of components should be enough to model all the contributions in each interval defined on the experimental data. Thus, by using narrow, automatically defined intervals it is expected that only few components are needed.

The number of models obtained depends on the number of intervals defined. Considering that for each minute of the chromatogram 85 intervals are defined, calculating eight models for each interval, a total of 680 models are fitted, corresponding to a total of 3060 components, for each minute of the chromatographic run.

5.1.3 Components selection

The component(s) selection step, compared to previous tools proposed with the same aim (Kamstrup-Nielsen, Johnsen, and Bro, 2013; Johnsen et al., 2014), is based on a

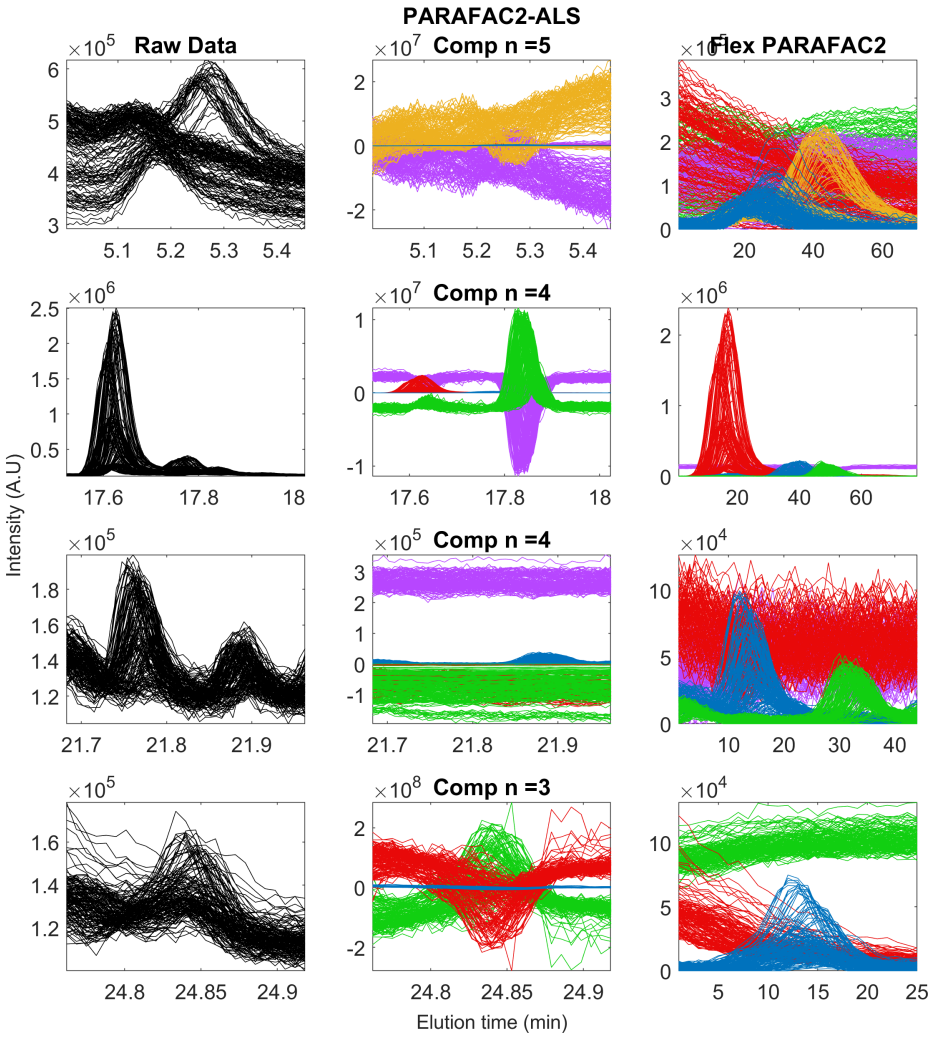

Figure 5.2 – Four intervals (black lines, first column) are shown. For each a PARAFAC2-ALS model (second column) and a flexible PARAFAC2 model (third column) have been fitted. In all the cases the models include the same number of components. It is possible to notice that the flexible PARAFAC2 models give a better solution compared to the PARAFAC2-ALS models.

different strategy. The idea of previous tools was to identify the appropriate model for each interval, and afterward the components describing the chemical compounds, i.e. the peaks, were extracted from the selected models to be included in the peak table. However, two considerations should be highlighted about this approach. The first point is that the identification of the right model for a given interval by means of these tools can be ambiguous and this aspect is highlighted also by the authors of the two articles where the tools have been proposed.

The second one is that the final selection of the components describing the peaks is left to the user, since the tools only identify the model with the right number of

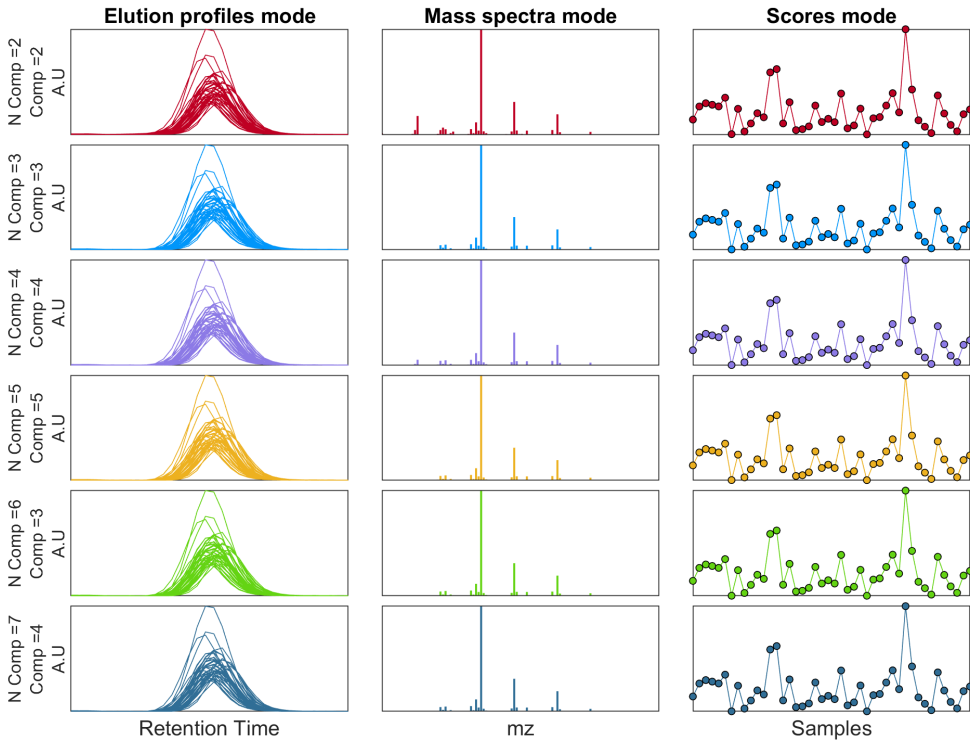

Figure 5.3 – Components extracted from seven different models, describing the same chemical compound.

components rather than selecting the components that are describing the peaks. Differently, AutoDise skips the preliminary step of the model identification and the peak table is obtained inspecting each and every model and selecting the components that are describing the peak. Skipping the preliminary selection of the models is based on the empirical observation that different models, i.e. models built on the same interval including different number of components or also models built on different intervals but covering the same peak, mostly describe a given contribution in the data equally well.

In figure 5.3 it is possible to notice that components extracted from different models built on the same interval can describe a specific contribution equally well; in table 5.1 the proposed identification, obtained comparing the resolved spectra with the NIST library, are shown. The identifications have been obtained considering the match factor, i.e. the inverse of the cosine distance between the unknown and the reference spectra (normalized between 0 and 999) (NIST, 2020).

Table 5.1 – Identified compound and match factor for the 6 components shown in figure 5.3.

Model	Component	Identified compound	Match factor
2	2	L-Alanine, ethyl ester	685
3	3	L-Alanine, ethyl ester	685
4	4	L-Alanine, ethyl ester	684
5	5	L-Alanine, ethyl ester	689
6	3	L-Alanine, ethyl ester	668
7	4	L-Alanine, ethyl ester	666

Thus, AutoDise is based on the analysis of the components rather than the analysis of the models. Each component from every PARAFAC2 model is analyzed in order to assess if it is describing a chemical compound.

5.1.4 Component screening

As mentioned before, each component from every PARAFAC2 model is analyzed in order to assess if it is describing a chemical compound, i.e. a peak. This is achieved with different tests. A first set of tests has been conceived to remove obvious artefacts:

1. The first of these tests is based on Tucker's congruence coefficient (Tucker, 1951; Lorenzo-Seva and Ten Berge, 2006) to identify and discard components affected by modeling artifacts, and it is conceived to reduce the number of false positives in the final peak table. Occasionally, models can include two or more of the modelled components with highly correlated trend; in particular, this happens with overfitted models. These components are not reflecting chemistry and this artifact often occurs as a problem related to the use of an improper number of components.

   Tucker's congruence coefficient measures the cosine of the angle between any pair of vectors, resulting in a range from -1 to 1; it can be seen as a non centered Pearson Correlation. The Tucker's congruence coefficient of the scores, of the mass spectra mode and of the mean profile of the elution mode of each component is calculated comparing the scores, the mass spectra mode and the mean profile

of the elution mode respectively of all the other components in the same model. When components included in the same models have the absolute value of at least one of the coefficients above 0.95, this set of components is discarded. This threshold has been adopted according to Lorenzo-Seva and Ten Berge, 2006, where it has been shown that components with a Tucker's congruence coefficient greater than this value are describing the same contribution.

2. Components that exceed the original signal intensity of the experimental data are filtered out as these indicate an invalid result. These components are not meaningful since the signals from significant components are supposed to be non negative and additive. Each component is analyzed and components exceeding the size of the corresponding data by a fraction higher to 10% are removed. This does not occur often and it would suggest some convergence problem during the fitting of the models.

This first screening has been conceived to remove components that are meaningless from a modeling point of view.
A second screening is subsequently performed to identify which of the remaining components are describing chemical compounds, i.e. the peaks. In this phase, a crucial role is played by the CNN network proposed by Risum and Bro (Risum and Bro, 2019). This is a deep neural network able to classify the elution mode profiles resolved by a PARAFAC2 model. It has been demonstrated that this model can successfully distinguish between four classes of elution profiles:

1. the label "Peak" is assigned to profiles describing a full elution peak;
2. the label "Shoulder" is assigned to profiles describing a cut-off or shoulder of a peak;
3. the label "Baseline" is assigned to flat profiles describing the baseline signal;
4. the label "Other" is assigned to mixed profiles that do not meet the shape of the other classes.

Another output of the CNN model is the probability-score, called "Niceness", for each and every resolved elution profile of each component. Basically, this means that for

each and every component, the model provides the probability of each resolved profiles of representing a peak. In this second step, the profiles resolved by the PARAFAC2 models are analyzed by the CNN and the classification results are applied to identify the components that are describing the peaks. It could be assumed that applying the model directly to the resolved profiles would result in a list of components to be included in the peak table. However, there are some considerations worth mentioning:

- as shown before, components from different models can describe the same peak. Thus, it is needed to identify the components that are describing the same chemical and select somehow only one representation for each peak.

- A second, important consideration is that it is not obvious how to use the classification results from the CNN model. A component includes as many profiles as the number of samples analysed, and it is not assumed that all samples have the same compounds. This means that not all profiles in a component describing a peak actually have the shape of a peak. The result is that the labels assigned to the profiles of a component by CNN are mixed and it is not always obvious whether a component is actually describing a peak or not. Therefore, criteria must be defined to exploit the classification results for component selection.

During the first developments of this step, the screening of the components was based only on the "Niceness". For each component, the samples with the highest "Niceness" were considered, in particular, the upper quartile of niceness. If this subset of profiles had a mean "Niceness" higher than .9 then the component was considered as describing peaks. However, during the development of AutoDise, and as a result of early applications on experimental metabolomics data, this single criterion proved insufficient to select components appropriately. It was common to include components with several profiles classified as "Shoulder" peaks, or with profiles characterized strange shapes classified as "Other".

In order to perform a more refined screening, other criteria have been conceived. During the development of these new criteria more attention has been paid on ensuring reliability than on the extraction of all available information, which explains that occasionally appropriate components are not selected.

As a result of this research, the parameters listed below were introduced. In particular,

the following parameters are assessed considering all the profiles belonging to a given component, and a component is assigned as a peak when:

1. the number of elution profiles classified as "Other" and "Shoulder" is lower than the number of profiles labeled as "Peak" and "Baseline".

   The profiles belonging to the first two classes are not important for the identification of the peaks, thus, if most of the profiles belongs to these classes, the component is discarded. Instead, if some of the estimated elution profiles of the specific PARAFAC2 component are describing a peak, and some of the profiles are labeled as "Baseline", it indicates that some of the analyzed samples do not include the specific compound thus the elution profile is flat resulting in a "Baseline" like shape.

2. The number of profiles labeled as "Shoulder" is lower than 10% of the total number of profiles belonging to the component.

   A high number of "Shoulder" profiles suggests that there is a peak, but the interval is misplaced with respect to it. Considering the high number of used intervals, it is likely that other models are centered on that peak, thus components describing shoulder peaks are discarded.

3. The mean peak probability score, over all profiles belonging to a given component, is greater than 0.3.

   This means that a component must include about the 30% of the profiles describing a peak; if the mean peak probability score is lower, the component is discarded.

4. The ratio of the variance of the profiles with a "Niceness" equal or higher than 0.6 to the variance of the profiles with a peak probability lower than 0.6 is smaller than a threshold of 0.4.

   This step indicates those components where only a few profiles, or samples, have been labeled as containing a peak, whereas most of them are not and with intense signals are removed.

Only components that meet all four criteria are selected as a peak. The result of this step is a list of components that are describing peaks. However another step is

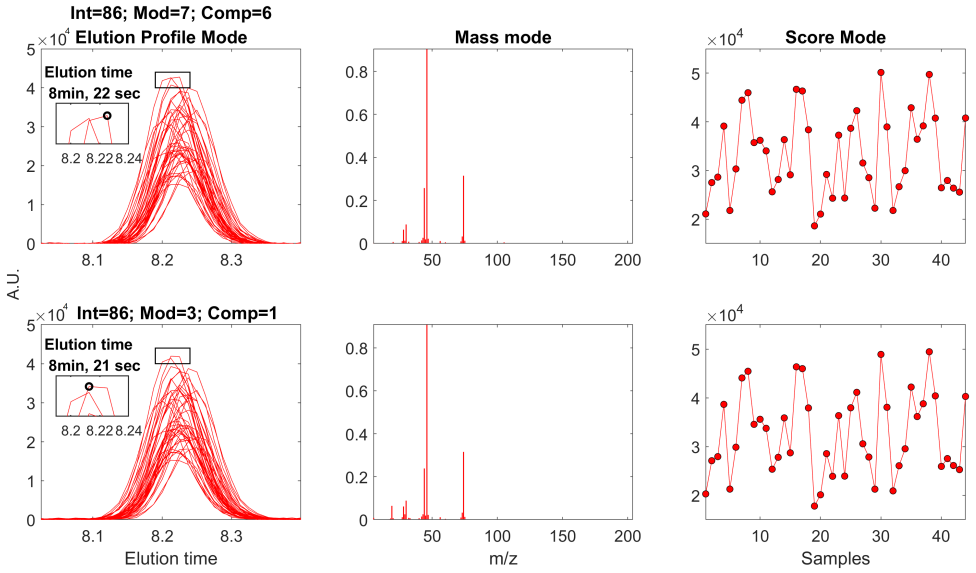

Figure 5.4 – Two components from two different models describing the same chemical compound, as it possible possible to notice from the mass mode and the score mode. It is possible to notice in the detail the shift of the maximum intensity

performed before to actually selecting the components to be included in the peak table. After the component screening and the elimination of those components which are not related to reliable peaks, the result is a list of components selected as describing peaks.

5.1.5 Component clustering

Considering how the intervals are defined, it is expected that several selected components represent the same chemical peak. In order to avoid including the same compound several times in the peak table, all the components with the same retention time are clustered, as different versions of the same compound are present and only the best of these several candidates is selected.

In the first versions of the algorithm, the clustering of the selected components considered only the elution time of the maximum intensity of the most intense profile. The assumption was that different components describing the same peak would have the maximum intensity at the same elution time. However, this assumption is not consistent with experience. For instance, it happens that components describing the same peak have the maximum intensity of the most intense profile at different elution times, as shown in figure 5.4

Moreover, another aspect that does not allow for consideration of only the maximum intensity is the retention time shift of the peaks. Because of the time shifts, the

peaks related to the same chemical across the samples have the maximum at different elution times. Considering how the intervals are defined in AutoDise, it happens that different (close) intervals do not include all the profiles related to a given peak. Thus, it can happen that the elution time of the maximum intensity of a given peak changes from component to component. If this is the case, considering only the retention time of the maximimum intensity, more components describing the same peak would be included in the peak table. To solve both these issues a less stringent clustering criteria has been applied.

The retention time of a component is defined as the average elution time of the maximum in the elution profiles mode for each component considering all the profiles. The value is rounded to the first digit (in minutes, using decimals rather than seconds). The result of this clustering is a set of groups including components from different models and/or intervals all describing the analytes eluting at the same time.

5.1.6 Clusters inspection

The rounding of the elution time for the cluster of the components may lead to more than one chemical compound being included, i.e. components describing co eluting peaks. Thus, in the final step, each cluster is analyzed to assess whether all the components grouped together are actually describing the same chemical compound. In particular, the resolved mass spectrum and concentration profiles of the components are compared by using the Tucker's congruence coefficient. If a given cluster comprises n components , two n-by-n matrices will be obtained, each containing Tucker's congruence coefficients for each couple of components. The idea is that components with a Tucker's coefficient higher than 0.8 for either or both the matrices are describing the same chemical compound. In this case, the threshold has been lowered compared to the proposed values in Lorenzo-Seva and Ten Berge, 2006, in order to set more stringent criteria to avoid redundancy in the final peak list, i.e. to avoid that a given peak is reported more than one time. When components within a group are describing different chemical compounds, they are divided in different subgroups according to the different chemical compounds. Thus, the components, previously grouped according to the elution time, are eventually divided into subgroups according to the chemical

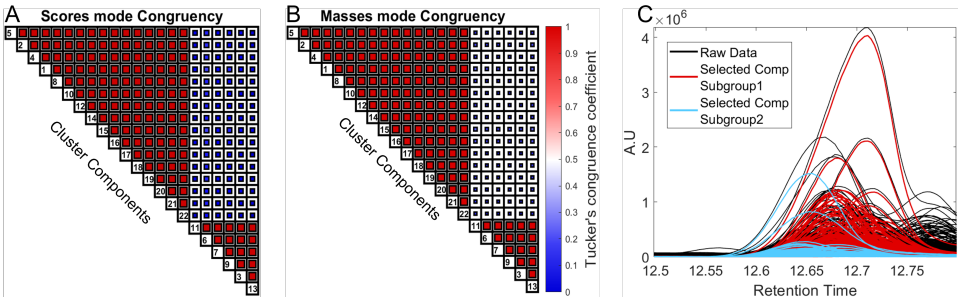

Figure 5.5 – A:Tucker's congruence coefficient plot of the score mode among the components included in the cluster. B:Tucker's congruence coefficient plot of the mass mode among the components included in the cluster. C: Components selected after the screening of the components within the cluster.

information extracted by the model. An example is given in figure 5.5. In the example a cluster made of 22 components is shown. From the Tucker's congruence coefficients calculated considering the resolved scores (figure 5.5 A) and masses (figure 5.5 B), it it possible to notice that the components are describing two different chemical compounds, resulting in two subgroups.

As a final filtering step, subgroups containing only a single component are discarded. One of the driving features of our system is the comparison and integration of components from a range of different models. When the system identified multiple components from different models describing the same information, i.e. peak and spectrum, this is a strong evidence for the correct identification of a peak underlying the data. If instead a component is found to not match any other components across all models, this indicates a less robust/trustable solution. These non-matching components are therefore removed from further analysis.

Finally, for each group/subgroup, the component with the highest "Niceness", as calculated by the deep neural network, is included in the peak table and used for the identification and relative quantification of the chemical constituent (figure 5.5 C). Other components within that group/subgroup are discarded or only used as insights to the grouping as part of the diagnostic tool.

5.1.7 Peak table definition

Once the components have been selected, the quantification of the corresponding compound for the analyzed samples is obtained multiplying the scores by the total norm of the elution profiles mode, where the norm of the elution profiles mode is determined as the norm of the relevant column of $\mathbf{B_k}$, resulting in the relative concentration across all the samples for the given molecule. The identification is performed by exporting the spectra of the selected components to the NIST library. Each resolved spectrum is compared to the library and the best identification is selected, based on the NIST match factor. The identification and quantification results, i.e. the peak table, are automatically saved in an .xlsx file called *autodise-table*. Details about the file are given in Appendix I.

It is important to clarify that the matching of the resolved mass spectra with any library should be considered as a first, tentative, identification of the given compound. Further identification of a molecule requires a cautious consideration of the experimental settings and most likely the analysis of standard compounds and/or the application of other analytical techniques such as NMR, in order to properly confirm the molecule identity and/or eventually correct and improve the result.

5.1.8 Case study: Analysis of GC-MS olive oils data.

One hundred sixty-two virgin olive oil samples, from extra virgin, virgin and lampante quality grades, have been analyzed by head-space solid-phase microextraction (HS-SPME) GC-MS. These samples are a subset of the samples used previously in (Quintanilla-Casas et al., 2020), where the details about the experimental settings are given. This could be considered a typical GC-MS dataset in terms of peak density and complexity (approximately 100-300 peaks), chromatogram length (around 58 minutes), acquisition rate (5.1 scans/s), sample size (162 samples), a visual representation of the TICs is given in figure 5.6).

The data are affected by severe retention time shift across the samples, this is due to a five-months analysis period. In untargeted analysis, quality control procedures are

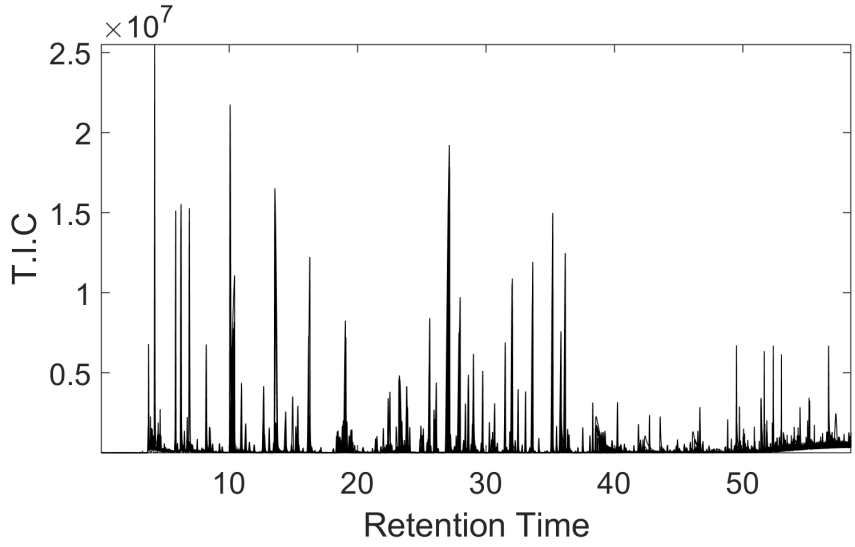

Figure 5.6 – TIC of the data.

a crucial aspect and several approaches have been proposed in this context, specifically when aiming for methods validation (Cavanna et al., 2018). However, issues such as peak shifts or variations in the analytical response are likely to appear and should be properly handled (Amigo, Skov, and Bro, 2010). Therefore, the GC-MS raw data set has been aligned to partially solve the peak shifts and improve the performance of the automatic interval selection, even though PARAFAC2 can deal with peak-shifts across the samples. This was done by means of a multistep approach: correlation optimized shifting (CoShift) of the TICs (Engelsen, Belton, and Jakobsen, 2005) followed by COW applied on the coshifted TIC and applied on individual m/z channels (Skov et al., 2006). These steps did not aim at a perfect sample alignment, the aim was to remove obvious misalignments and thus improve the performance of the automatic interval selection, even though PARAFAC2 can model peak shifts across the samples. The data have been split in six slightly overlapping calculation-batches considering the elution time axis and sequentially analyzed one at time to avoid computer memory problems during the calculation of the models and to limit the computational cost of each round. This was due to the dimension of the dataset and the number of intervals defined.

A total of 374,184 components have been calculated. Overall, the calculation of the models took two weeks, the screening of all the components required less than one hour. The models have been calculated with a Intel® Core™ i7-6950X CPU processor with

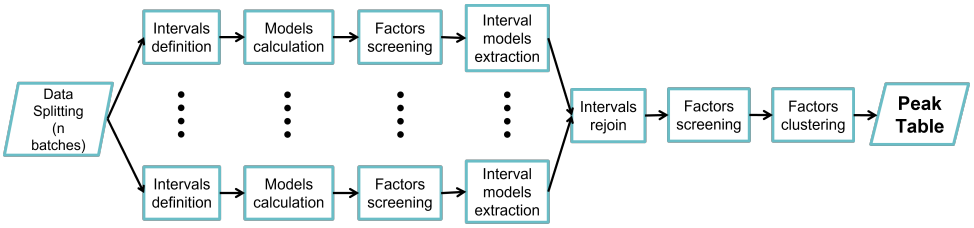

Figure 5.7 – Schematic representation of the AutoDise workflow applied for the case study. The data have been split into batches considering the time axis. For each batch, the intervals have been automatically defined, the models have been fitted looping over all the intervals. The selection of the components took place once all the models have been calculated. When AutoDise detected a component as describing a chemical compound, all the models calculated considering the corresponding interval have been extracted. The selected models, from all the batches, have been joined together and a final round of component screening, clustering and selection have been performed to obtain the peak table.

a dedicated RAM of 32 GigaBytes. The calculated models have a final dimension in the order of 70 GigaBytes, about 10 GigaBytes for each batch. Because of the split of the data in batches, slight modifications to the original AutoDise algorithm have been adopted, as shown in figure 5.7. At first, AutoDise was applied on the models of each batch, to keep only the relevant information within each of them. In particular, all the models for a given interval were extracted when at least one component of the models was selected. In total, 328 intervals have been extracted from the six batches. AutoDise was applied a second time considering the 328 intervals to avoid edge artifacts between the batches. As a result, a final peak list including 340 chemical compounds has been obtained. The relative concentration and the tentative identification based on the NIST matching have been obtained for each compound.

As a first assessment, the performance of AutoDise has been visually evaluated, assessing if the proposed solutions are adequate. All the selected components account for a corresponding peak in a meaningful way (figure 5.8 A). However, the absence of components for some prominent peaks can be noted (figure 5.8 B, C, D). Although figure 5.8 gives an overview of AutoDise performance, a thorough assessment has been performed, comparing the automated results with a comprehensive manual analysis.

This manual analysis resulted in a total of 133 identified and quantified compounds. Manual tentative identification, integration and semi-quantification steps were performed by average-experienced GC-MS users by means of the ChemStation (Agilent)

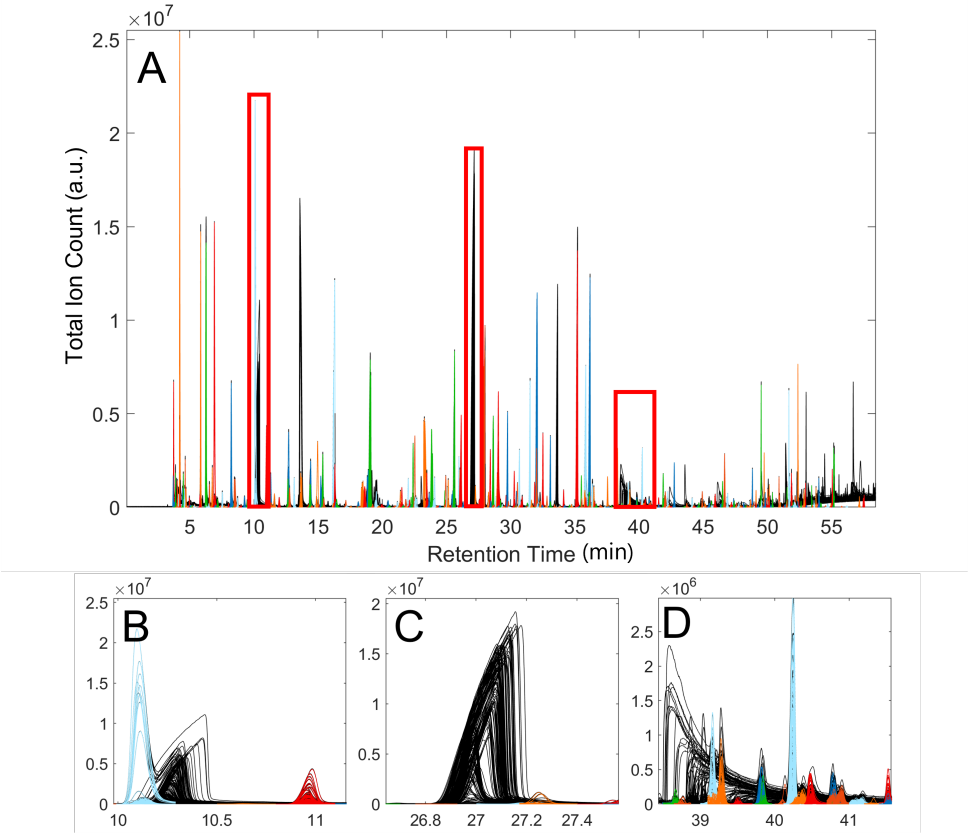

Figure 5.8 – A: The TIC of the samples is shown in black. The components selected by the expert system are shown and represented by the coloured lines. The differences in colour serve to aid visualization. B, C, D: three details of some problematic peaks discussed in the text. The areas shown correspond to the three red boxes in 5.8 A. Reproduced from (Baccolo et al., 2021)

software. Compounds were tentatively identified comparing the corresponding mass spectra and retention times with those available in the WILEY mass spectrum library. The number of compounds (n=340) identified by AutoDise is more than twice compared to the number obtained from the manual procedure (n=133). Thanks to the exhaustive peak table obtained by AutoDise, several of the compounds omitted or not detected in the manual analysis could be indeed linked to oxidative, fermentative and biogenetic pathways and, thus, relevant for the quality of virgin olive oils, thus providing more information compared to the manual analysis

The peak identities and retention times have been used for the comparison of the two peak tables. In the peak table obtained by AutoDise, 118 out of 133 peaks (89%) identified with manual analysis using ChemStation were included. The Pearson correlation coefficients of the relative concentrations across all 162 samples as provided

Table 5.2 – Pearson correlation coefficients among the concentrations estimated by AutoDise and complementary manual PARADISe and concentrations estimated by the manual analysis. Reproduced from (Baccolo et al., 2021).

Target compounds (n=133)		
Pearson Correlation coefficient	AutoDise	PARADISe
>0.9	99	12
0.7-0.9	10	2
<0.7	5	1
NA	4	-
TOTAL	118	15

by PARADISe and the manual analysis were calculated and summarized in table 5.2. The 84 % of these compounds (99/118) had a correlation coefficient above 0.9. The remaining 13% of the compounds (15/118) had a correlation coefficients below 0.9. Analyzing all the specific cases, it has been realized that errors were due to the incorrect integration during the manual analysis. Many of these compounds were present at low levels and generally coeluted with more abundant compounds, complicating both the manual identification and the manual quantification steps. For the 3% of the compounds (4/118), it was not possible to calculate the correlation between the concentrations estimated by AutoDise with the manual analysis as the latter was estimated including coeluting compounds.

In order to further analyze the AutoDise results, a thorough investigation of the fifteen compounds identified by the manual analysis and not detected by the automatic analysis has been performed. Overall three distinct issues have been found:

Interval selection issues In this case the intervals automatically defined were not wide enough to include very intense peaks, such as those represented in figure 5.8 B,C,D. Twelve of the fifteen were not included because of limited interval range.

Component selection issues The remaining compounds were excluded by Autodise during the analysis of the components despite the fact that components for these peaks were obtained. An example is the peak corresponding to 2-methylbutyl acetate. Manually investigating the models obtained by AutoDise, a component describing the chemical compound has been found (figure 5.9 A), but the component which explained this chemical compound was not reported by the automated system.

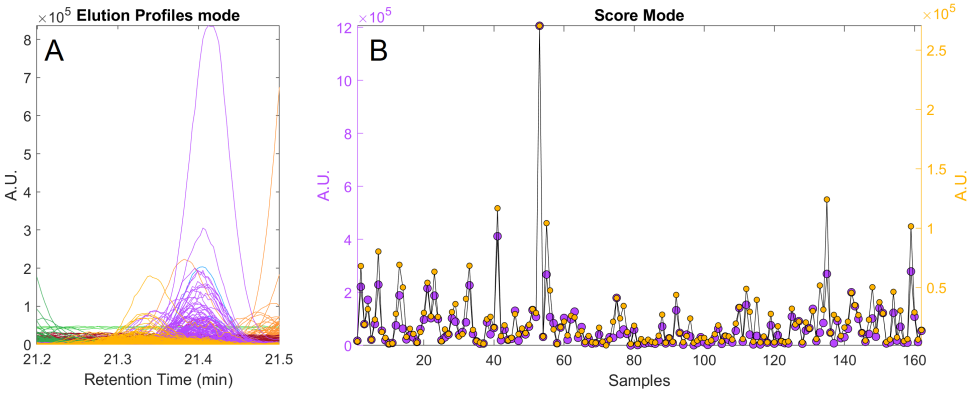

Figure 5.9 – A: elution profile modes of the model. The component corresponding to 2-methylbutyl acetate is colored in yellow, the component profiles with high Tucker's congruency are colored in purple. B: Score modes of the two components, the two profiles almost entirely overlap each other. The figure has a double axis so to better appreciate the coherent trend between the two components. Reproduced from (Baccolo et al., 2021).

The reason why is due to the high Tucker's congruency of the scores mode with another component within the model (figure 5.9B).

As described in the previous section, during the development of the criteria for the components selection greater attention was focused on ensuring reliable solutions than on exploiting all available information, and this is the reason why appropriate components were occasionally not selected. For the two cases above, issues related to the interval definition step can be easily solved by manually defining an interval that spans the full peak, then AutoDise can be applied as well on the new models fitted on the manual selected intervals to select the best solutions. This strategy has been adopted for 12 of the 15 above mentioned missing compounds, where no appropriate interval had been defined automatically (table 5.2).

The solution of issues related to the second case outlined above is more laborious. The user should manually analyze and select the component that corresponds to the observed peak, as routinely performed with PARADISe.

As a result, comparing the manual targeted approach and the manual PARADISe, 80% of compounds achieved a good correlation (> 0.9). The summarized results are available in table 5.2, and an extension of this table can be found at Baccolo et al., 2021. The offered solutions for the fifteen missed compounds show that, although automatic selection system performed very well compared to a fully manual approach, there is still room for improvement.

An experienced GC-MS user could still play role in data treatment process, since knowledge about what to expect from the analysis of a given sample set is essential to check and interpret the resulting peak table. As such, AutoDise should not be taken as a black box. It is conceived to ease the analysis of GC-MS data, but the analysis and the interpretation of the results is a crucial aspect that should be performed by skilled analysts. In particular, it would be necessary to check when an important peak is missed and, if so, to be able to manually run PARADISe or an alternative method. Despite this, the benefits that the automated PARADISe provides regarding the GC-MS data treatment are noteworthy, both in terms of time saving, tentative identification and relative quantification accuracy and comprehensiveness of the chromatographic data.

5.1.9 AutoDise GUI

In this section, the GUI (MATLAB 2021a) developed to implement the AutoDise algorithm is described.

It could be argued that the description of a GUI is not relevant in a scientific publication. However at the same time, when new findings are achieved in machine learning and chemometrics, they often appear in terms of a mathematical protocol (an algorithm), written in a scientific paper. This is of great value in itself, but will only be accessible to people that understand that language, so to speak. Nevertheless, the algorithmic developments presented in this thesis, while interesting to the chemometrics community, also has direct and immediate relevance for a scientific audience that does not speak our scientific language.

In order to make these findings accessible to more people, that is, in order to assure our chemometrics findings have impact and benefit to a wider scientific audience and also to the society, we need to do more than just publish algorithms. At the lowest level we can make software implementations in MATLAB, R, Python etc. but again, not all people are able to use such programs. Therefore, taking things one step further and making graphical user interfaces for our algorithms is scientifically needed, in order for our results to gain its intended use.

The GUI is organized in four tabs; the functionalities of each tab are described in this section.

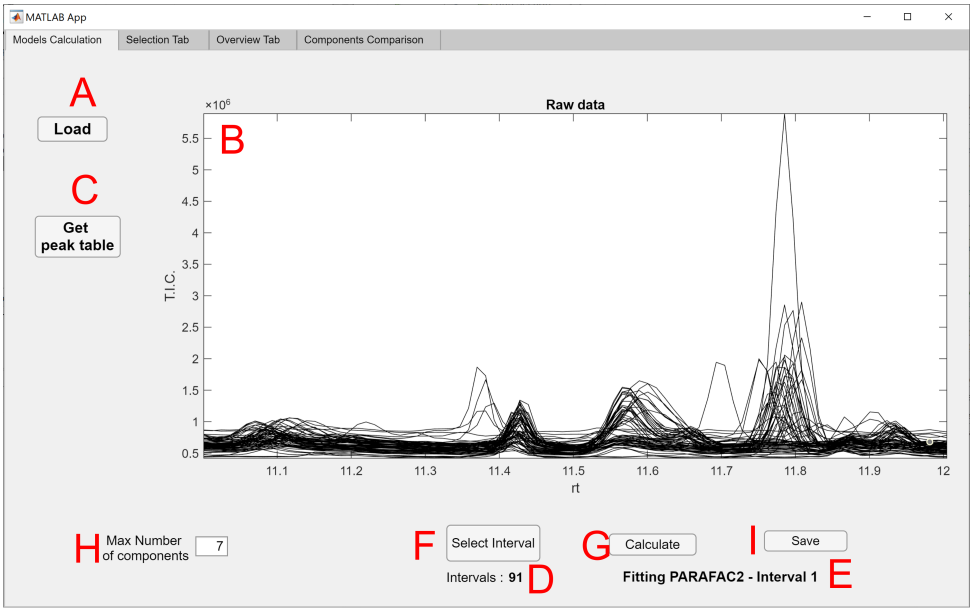

Figure 5.10 – First tab of the AutoDise GUI.

Tab 1: Models calculation

The first tab, called *Models Calculation*, of the AutoDise GUI is shown in figure 5.10. The **Load** (figure 5.10 A) button allows to load the data. The data must be a .mat file including a structure called *Data*. The structure must include the following fields:

M : in this field the three way GC-MS data are stored,

- the first dimension corresponds to the samples;
- the second dimension corresponds to the elution time;
- the third dimension corresponds to the mass spectra.

TIC: in this field the two way TIC of the GC-MS data are stored,

- the first dimension corresponds to the samples;
- the second dimension corresponds to the elution time;

rt: in this field the vector holding the elution time values is stored. Both a column or row vector format is accepted.

mz: in this field the vector holding the mass spectra range is stored. Both column or row vector format is accepted.

FileNames: in this field is stored a cell vector holding the name of the samples. Each cell must contain a single name. Only column vector is accepted.

Box – Tip

If only the .cdf files are available, they can be loaded in PARADISe v5.X; once loaded the data can be saved as .mat. The file generated by PARADISe is accepted by the AutoDise GUI.

Once the data are loaded, the TIC of the data are automatically plotted in the main window (figure 5.10 B). To start the analysis of the data, the user has two options:

- the full auto analysis;
- the semi auto analysis.

The first option is selected by clicking the **Get peak table** button (figure 5.10 C). Selecting this option, the loaded data are passed through the whole AutoDise workflow as described in the previous section. The number of intervals is printed on the tab (figure 5.10 D). To check if the calculation is proceeding, the number of the interval currently being processed is printed on the tab (figure 5.10 E).

Box – Tip

It is a good practice to not calculate too many models in a single run. If more than 1000 intervals are defined, the suggestion is to split the data in different batches as in the previous section.

Once all the models are calculated, the components selection step is perfomed. Once the selection is completed, a window will appear to select the folder where to save the peak table. The peak table is saved in a .xlsx file named *autodise-table*. Five sheets are included in the file.

Overview: in the first sheet information about the analyzed data, the analysis pipeline and a legend of the other sheets is given.

Relative Concentrations: the second sheet contains the relative concentrations of each compound across all the samples.

Resolved mass spectra: in the third sheet the resolved mass spectra for all the selected compounds are reported.

Top NIST hits: in the fourth sheet the top two NIST hits for each compound are reported.

Interval Details: the fifth sheet contains information and diagnostics about the models from which the components have been extracted.

More details about the file are given in Appendix I.

The difference between the full auto analysis and the semi auto analysis is that in the second option, the intervals are manually defined by the user and not automatically. To manually define the intervals the user has to click the *Select Interval* button (figure 5.10 F). A cursor will appear and the boundaries of the interval can be defined by clicking the start point and the end point on the TIC plot (figure 5.10 B).

Box – Tip

All the axes in the AutoDise GUI can be resized. Zooming in the TIC before manually selecting the interval boundaries can be of great help.

The button must be pressed for each interval, thus after the limits are defined for an interval, if the user wants to select another one the button must be clicked another time. Each time an interval is defined, the interval counter (figure 5.10 D) will be updated. Once all the intervals have been defined, pushing the *Calculate* button, the calculation of the models will begin. The progression of the calculation can be monitored looking at the message printed on the tab (5.10 E). Once the calculation of the models and the selection of the components are terminated a window will appear to select the folder where to save the peak table as an .xlsx file called *autodise-table*.

Box – Tip

The only parameter that can be set by the user is the maximum number of components included in the models. By default, it is set to seven (most of the time it is enough to extract all the chemical compounds), but it can be changed

in the box shown in figure 5.10 H. The higher the number, a longer time for the calculation of the models.

Both in the case of full auto and semi auto analysis, the data and the calculated models can be saved by pressing the **Save** button (figure 5.10 I). A window will appear to select the folder where to save them as single .mat file called AutoDise_Session. The file can be loaded in the GUI using the load button in the second tab, described below.

Tab 2: Selection tab

The second tab of the AutoDIse GUI is called Selection tab, shown in figure 5.11. This tab can be used to easily evaluate the choices and solutions given by AutoDise and verify if the selected components are good or not. In this tab, it is possible to load previous sessions using the **Load** button (figure 5.11 A), moreover the GUI can also load sessions from the PARADISe platform v5.X.

Once the data are loaded, the number of intervals and the total number of components loaded are updated below the **Load** button. Clicking the **Selection** button (figure 5.11 B), the models are analyzed and the components selected, as described in the previous section. Once the selection is completed (the status is printed on the tab (figure 5.11 C), it is possible to select the second subtab, called *Results analysis*.

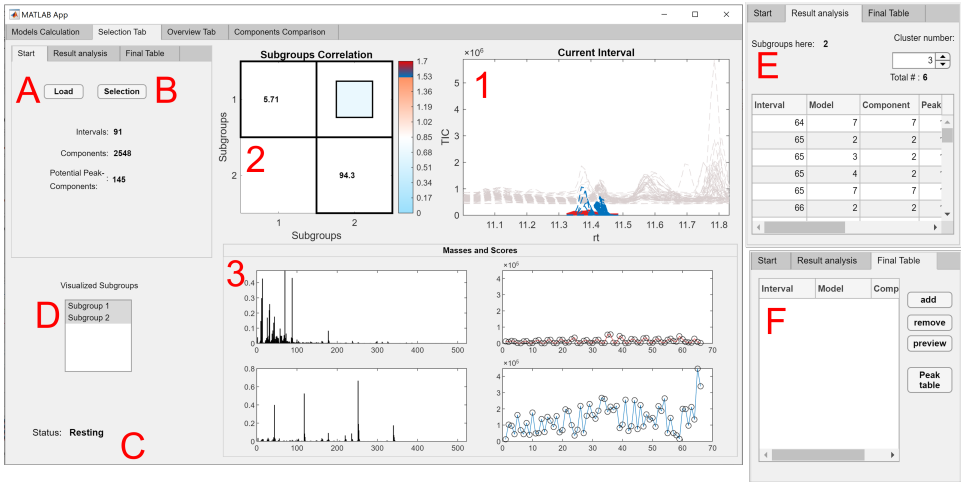

Figure 5.11 – Second tab of the AutoDise GUI.

Box – Tip

It is possible to check the results immediately after they have been obtained. After the peak table has been saved it is possible to switch to this tab and start the analysis, the data and the models are already loaded.

In this subtab, shown in figure 5.11 E, it is possible to explore the different clusters and subgroups obtained during the selection of the components. Using the *Cluster* spinner, it is possible to explore the different clusters. For each cluster:

- the number of subgroups for the specific cluster is reported (in the example shown in figure 5.11 the number of subgroups within the cluster is 2);
- each component included in the cluster is reported in the table. The components are identified by the interval number, the model from which they were extracted (the model is identified by the total number of components included), and the specific component number. In the table, the subgroup and so-called niceness for each component is also indicated. The parameter Niceness is used to select the best component for each cluster or subgroup.

The plots 1 2 and 3 in figure 5.11 are automatically updated when a new cluster is selected.

In the **Current Interval** axes (figure 5.11 1) the selected component(s) (coloured) over-impressed on the raw data (grey) are shown.
In the **Masses and Scores** panel the mass and the score modes for the selected component(s) from each cluster/subgroup are shown. Each row corresponds to a given subgroup and the colors are coherent with those in 5.11 1. An example is shown in figure 5.11 3. A maximum of four subgroups can be visualized in the **Masses and Scores** panel. If more than four subgroups are present for a given cluster, it is possible to specify the subgroups to plot in the box (figure 5.11 D).

The plot in figure 5.11 2 is shown only for the clusters where subgroups have been detected. It shows the subgroups correlation in terms of Tucker's congruence

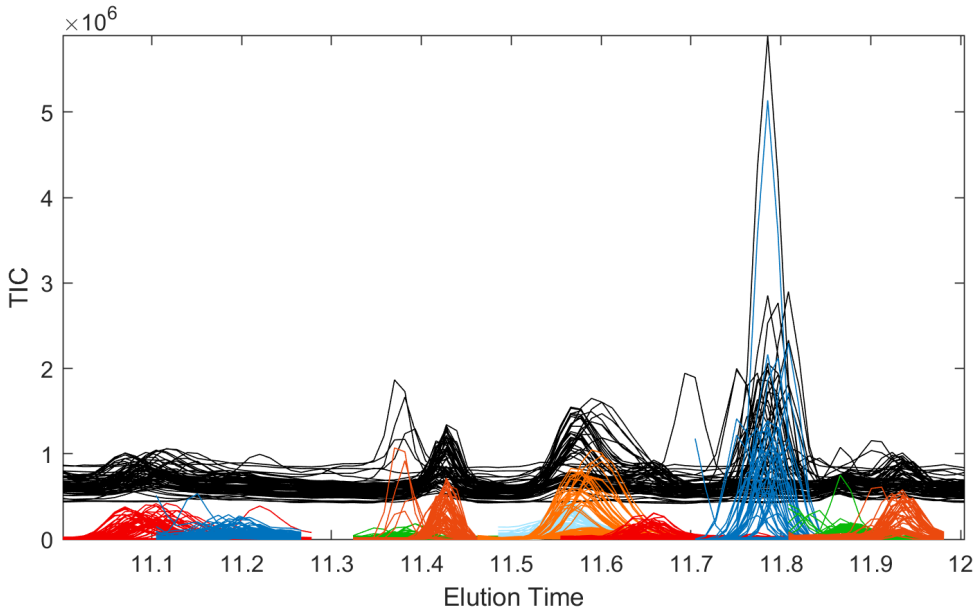

Figure 5.12 – Preview plot showing the selected components (colored) overlaid on the raw data (black).

coefficients, as explained in the previous section. It consists in the sum of the cosine distance between the spectra mode and the correlation mode between the best components selected for each subgroup (colored square) and how the components are distributed across the subgroup in terms of percentage. In the example, 5.71% and 94.3% components belong to the subgroup 1 and subgroup 2 respectively. The correlation across the best components of the two subgroups is about 0.7, indicating that the two components are describing different chemicals.

In the third subtab, called *Final table* (5.11 F), it is possible to add unselected components to the Peak table. Pressing the **add** button, a new row will appear in the table and specifying the interval number, the model from where to extract the component (identified by the total number of components included) and the actual component number, the component will be included in the peak table. To remove an added component, the **remove** button can be clicked and the last row of the table is deleted. The **preview** button produces a plot, where all the selected components are superimposed on the raw data. An example is given in figure 5.12.

Pressing the **Peak table** button a window will appear to select the folder where the peak table is to be saved as an .xlsx file, with the same structure as described before.

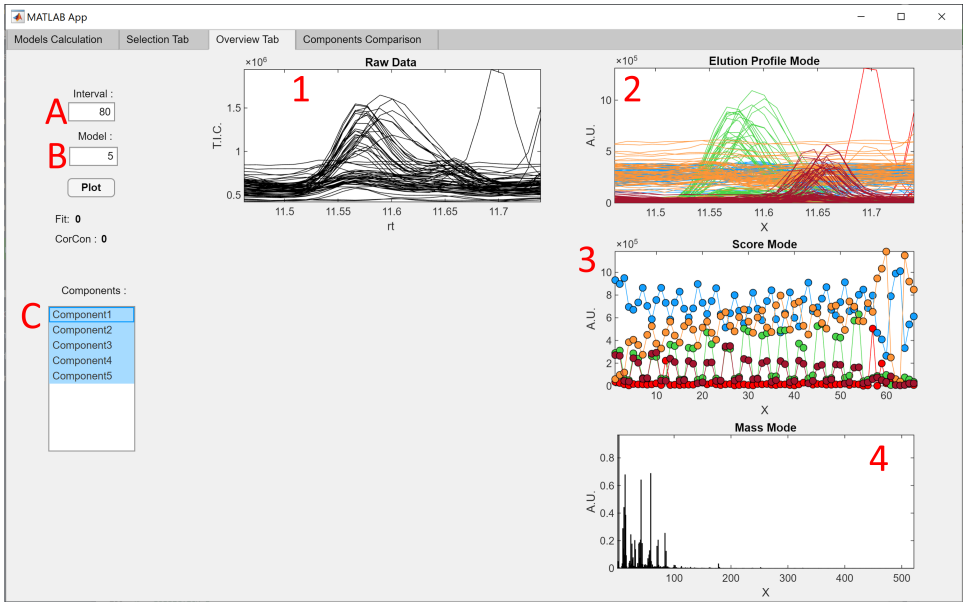

Figure 5.13 – Third tab of the AutoDise GUI.

Tab 3: Overview tab

In the *Overview* tab, shown in figure 5.13, it is possible to have an overview of a specific model. To plot a model, the interval and the model, defined by the total number of components, must be indicated in the two boxes in the tab (A, B in figure 5.13). Once the two boxes have been defined clicking on the **Plot** button, the raw data, the elution profiles mode, the scores and the mass mode of all the components included in the model are plotted in the graphs 1, 2, 3, and 4, as indicated in figure 5.13. In the *Components* box it is possible to select specific component(s) to be plotted. The plots in the graphs 1, 2, 3, and 4 are automatically updated, showing the selected component(s).

Components comparison

The *Components Comparison* tab is shown in figure 5.14. It can be used to compare a maximum of three single components. To plot a component the user has to define in the three boxes (A, B and C in figure 5.14) the interval number, the model from which the component is selected (identified by the total number of components included in the model) and the actual component number to be plotted, respectively. Clicking on the **Plot** button the indicated component is plotted. In particular the axes 1, 2 and 3 highlighted in figure 5.14 will correspond to the elution profile mode, the score mode

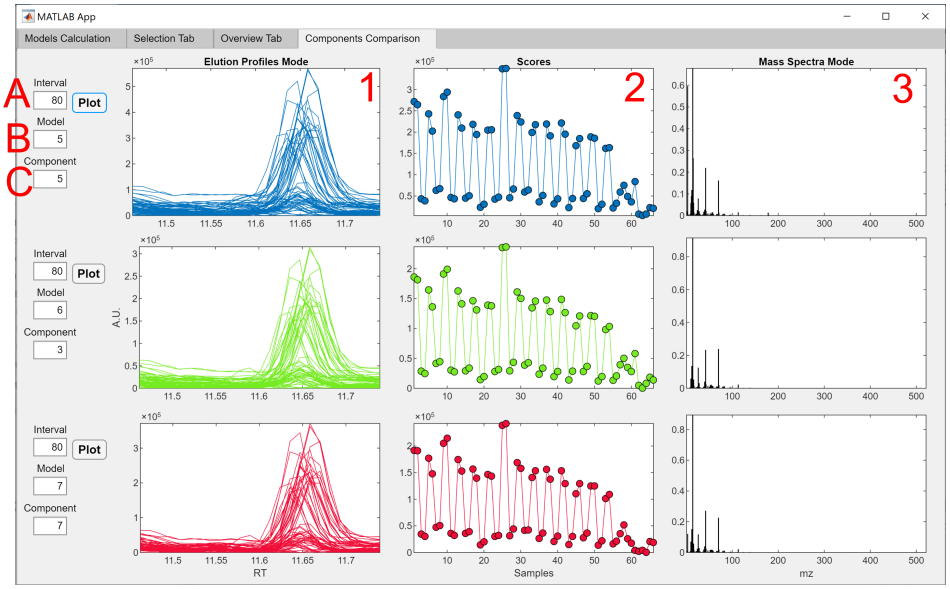

Figure 5.14 – Fourth tab of the AutoDise GUI.

and the mass spectra mode, respectively. In the example shown in figure 5.14, three components (5, 3 and 7) from models including 5, 6 and 7 components respectively, built on the interval 80, are shown.

5.1.10 Workflow optimization

The original workflow for AutoDise is described in the previous section and reported in figure 5.7.

In the case study reported in the previous section, the data have been split in six different batches and analyzed one at time to avoid computer memory problems during the calculation of the models and to limit the computational cost of each round. This was due to the dimension of the dataset and the number of intervals defined. In the case study reported in the previous section, a total of 374,184 components have been calculated.

However, from the workflow in 5.7 it is possible to notice that some of the operations, such as the component screening, are repeated several times and that all the components are saved during the calculation of the models. Considering all the six batches, the generated files were in order of 12 Gigabytes each, for a total of about 96 Gigabytes. Moreover, out of the 374,184 calculated components, only 1885 were selected by AutoDise, corresponding to 0.5% of the total. In order to reduce the computational

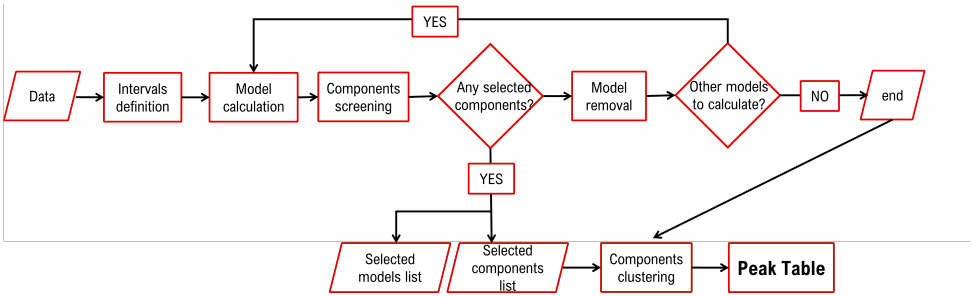

Figure 5.15 – New designed workflow for AutoDise

cost and the memory requirements, a new worflow, schematized in figure 5.15, has been designed. The main differences compared to the previous workflow are:

1. the data do not need to be split;
2. the component selection is performed immediately after a model has been calculated;
3. if no component(s) have been selected, the model is deleted;
4. once all the models are calculated, the clustering step is performed only considering the selected components;

Keeping only the models with selected components will drastically reduce the memory needed to store the results. The new workflow will be implemented in the next version of PARADISe (v6.0).

5.2 Machine Learning approaches for elution profiles classification

As described in the previous section, the CNN proposed in Risum and Bro, 2019 plays a crucial role in AutoDise to identify the components describing the peaks, i.e. a chemical compound. However, despite its effectiveness, during the optimization of the system, some issues in the identification of the components have been noticed related to the performances of the deep neural network, i.e. some elution profiles of components describing peaks were not labelled as "Peak" by the CNN and thus not included in the peak table. Investigating the reasons behind and looking for solutions, the main source of ambiguity for the model was found in the original data used to train the model. In particular, by visually inspecting the original data used during the training, it has

been noticed that profiles with similar shapes were labeled as belonging to different classes. This source of ambiguity cannot be handled by the trained model, thus a new dataset made of manually labeled resolved elution profiles has been prepared and used to train a new CNN model with the same architecture proposed in Risum and Bro, 2019, as well as CNN, a BILSTM model and a KNN model used as a benchmark.
The three approaches have been selected specifically considering the task and the features of the samples. A kNN model classifies a new sample according to the most frequent class of the nearest samples. The quality of the model is expected to increase with the number of training samples. Considering the huge number of profiles generated for this work, it appeared a viable alternative for this task. Convolutional neural networks identify patterns within a sample regardless of the specific position and also the performances shown in Risum and Bro, 2019 support the effectiveness of this architecture, handling the retention time shifts. RNN, which LSTM and BILSTM networks are part of, have been widely applied for classification of time series data, and elution profiles belong to this category.

5.2.1 Dataset preparation

Head-space solid-phase microextraction (HS-SPME) GC-MS data obtained analyzing 66 olive oil samples were used to obtain resolved elution profiles by means of PARAFAC2. The experimental details are available at Quintanilla-Casas et al., 2020. A total of 44 intervals on the time dimension have been defined and all the intervals were resolved by PARAFAC2 modelling. A total of 306 PARAFAC2 models were calculated, resulting in 1214 components. A total of 80124 estimated elution profiles were thus obtained (1214 components x 66 samples) and then used as input of the classification models. In particular, the 80124 profiles were randomly split into two sets: 68106 profiles (85%) were included in the training set and the remaining 12018 profiles (15%) in the internal validation set. The models were trained on the training set and their hyperparameters were tuned based on the predictions of the internal validation set profiles. Moreover, an external test set has been prepared to evaluate the predictive ability of the trained models. It contains 7673 profiles preprocessed with PARAFAC2, as the training data. These profiles have been retrieved as a subset from the test set in Risum and Bro, 2019. To increase the number and the variability of the data, crucial requirements for

all the tested models, all the profiles have been duplicated and horizontally flipped, resulting in a total of 175594 profiles.

5.2.2 Labelling

All the profiles have been manually labeled according to four classes: 'Peak', 'Cutoff peak', 'Baseline' and 'Other'. A preprocessing routine, according to Risum and Bro, 2019 was applied to all the profiles. First, each profile was linearly interpolated to a length of 50 and normalized to unit vector, i.e., a vector with a norm equal to one. This preprocessing equals the size of the elution profiles and, thus, allows the classification of profiles with different lengths. At the same time, the normalization maximizes the comparability of the profiles.

After the normalization, all the profiles were visually inspected and manually labeled. A set of semi quantitative criteria was defined to be consistent during the manual labelling of the elution profiles.

1. For a profile that is labelled as "Peak", the part of the time interval where the peak is present is visually defined. The remaining part is called tails. The maximum of the peak has to be higher than 0.1, the maximum of the tails lower than 0.1, otherwise the profile is labelled as "Cutoff peak" or "Other' depending on the specific profile (figure 5.16 1 B, C).

2. For a profile that is labelled as "Cut-off peak", the part of the time interval where the cut off peak is present is visually defined. The remaining part is called tail. The maximum of the cut-off peak has to be higher than 0.1, the cut-off peak has a minimum value to the right (Figure 5.16 B) or the left of the maximum (Figure 5.16 A) higher than 0.1, the maximum of the tail lower than 0.1, otherwise the profile is labelled as "Peak" or "Other" depending on the specific profile (figure 5.16 B, C).

3. For a profile that is labelled as "Baseline" the flat or monotonically increasing or decreasing trend is visually defined. The difference between the maximum and minimum signal value has to be smaller than 0.2 otherwise the profile is labelled

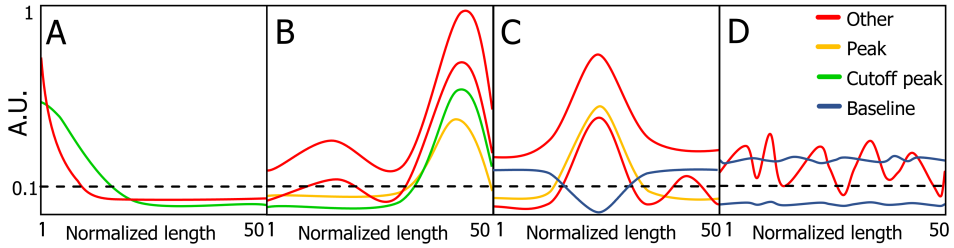

Figure 5.16 – Graphical representation of the labeling criteria. The dotted line highlights the 0.1 threshold.

as 'Other' (Figure 5.16 D). Baseline profiles may show a negative peak (figure 5.16 1 C).

4. A profile is labelled as "Other" if it does not meet the criteria for the above classes.

It should be noted that the set of rules does not include any indication about the overall shape for any of the considered classes, and also the noise is not considered. This is important for instance in the class "Baseline", where a strict monotonic trend is extremely unlikely, because of the noise. Moreover, profiles showing a not negligible amount of noise have been labelled as "Other". Nonetheless, a given threshold to discriminate profiles based on noise has not been defined and the choice to assign a profile to the class "Other" or to the most appropriate class was made according to the visual evaluation of each profile.

As such, the proposed criteria can be seen as a reasonable rule of thumb to prevent contradictions during the labelling phase, but cannot be automatically applied for the labelling of the profiles. In total, including the doubling of the profiles, 175594 profiles have been manually labelled and used during the training, the validation and the test of the models. Details about the class distribution and the number of profiles for each set are shown in table 5.3.

5.2.3 Models

5.2.4 CNN

Convolutional Neural Networks (CNN) are a family of neural networks widely applied in image recognition. Several different variations in CNN architectures have been

Table 5.3 – Number and distribution of profiles included in the training, validation, and test sets.

	Other	Peak	Cutoff peak	Baseline	Total
Training set	30.5% 41528	18.9% 25794	6.6% 8932	44% 59958	136212
Validation set	29.7% 7150	19.5% 4698	6.4% 1548	44.3% 10640	24036
Test set	34.5% 5296	24% 3682	15.1% 2312	26.4% 4056	15346

proposed, but in general they consist of stacked convolutional and pooling layers, followed by one or more fully connected layer(s). The convolutional layer is the core of a CNN, and it is based on a set of trainable filters or kernels. Details about the theory of CNN are given in chapter 3.

The architecture of the CNN model has been implemented as described by Risum and Bro, 2019. The network is made of four convolutional layers followed by two dense layers. No further optimization was performed. Details about the settings and optimization can be found in Risum and Bro, 2019.

5.2.5 (BI)LSTM

Recurrent neural networks (RNN) are a family of neural networks used to deal with sequential data and to capture long-range dependencies between sequence data (Rumelhart, Hinton, and Williams, 1986). In particular, RNNs with Long short-term memory (LSTM) and Bi Linear LSTM (BILSTM) units were considered (Hochreiter and Schmidhuber, 1997). Details about RNN and LSTM and BILSTM are given in chapter 4. In speech recognition, where (BI)LSTM networks are widely used, long-range dependencies are important since the meaning of a sentence changes depending on how the words are arranged. Hence, keeping track of the positions of the words, even when they are not immediately next to each other, is crucial to understand the sense of the phrase (Sundermeyer, Schlüter, and Ney, 2012). The same concept can be applied for the classification of elution profiles. For instance, let consider a "Peak" and a "Cutoff peak". In general, these profiles are both characterized by the same patterns, e.g., peak or flat curves. The difference between a peak and a cutoff peak is how these

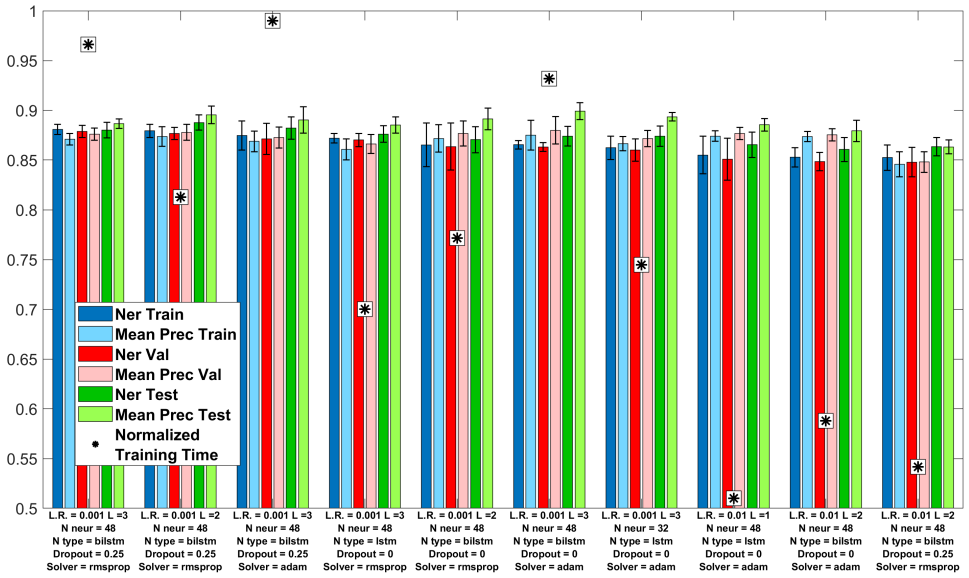

Figure 5.17 – Results of the five replicates for the 10 best BILSTM settings. The bars represent the standard deviations. L.R stands for learning rate, NER for non error rate, N neur for number of neurons in the first hidden layer

patterns are placed through the profile. Therefore, it is important to "remember" how the different, and even distant, parts of a profile are organized. The (BI)LSTM networks are expected to handle this kind of task.

The general architecture for all the recurrent models tested during the optimization included the following layers: an input layer, (BI)LSTM layer(s), a fully connected layer and an output layer. This architecture implies the tuning of the hyperparameters listed in table 5.4. In particular, two values of initial learning rate were considered (0.001 and 0.01). The input layer is made of 50 neurons (i.e., the length of the normalized profiles), and this was kept constant across all the tested networks. One, two or three stacked layers of (BI)LSTM followed by one fully connected layer have been tested. The same layer type has been used for each layer, therefore no combinations of LSTM and BILSTM layers have been tested in multiple layer models.

The number of neurons in the first layer was set to 16 or 32 or 48, and the number of neurons was halved each time for each successive (BI)LSTM layer, i.e., the number of neurons for the second and third layer was half and a fourth of the neurons in the first layer, respectively. To avoid overfitting, a dropout of 0.25 (Srivastava et al., 2014) as a regularization term for the first layer has been introduced, but unregularized

Table 5.4 – Parameters tested and selected for the BILSTM network.

Parameters	Option 1	Option 2	Option 3	Selected Option
Initial learning rate	0.001	0.01		0.001
Neurons Type	LSTM	BILSTM		BILSTM
Dropout	0	0.25		0.25
Solver	Adam	RMSprop		RMSprop
Activation function	ReLu	tanh	sigmoid	tanh
Num of (BI)LSTM layers	1	2	3	3
Num of neurons in the first layer	16	32	48	48

networks were considered (i.e., dropout equal to 0).

The number of neurons in the fully connected layer is the same as the last (BI)LSTM layer. Three different activation functions have been considered: ReLU, sigmoid and hyperbolic tangent. The output layer with a softmax function (four neurons, one for each class) has been kept constant across all the models. Two optimization algorithms or solvers have been tried: Adam and RMSProp (Ruder, 2016). A preliminary optimization of the hyperparameters has been performed by means of a full grid search considering the combination of all the parameters listed in table 5.4, for a total of 432 tested networks. The ten architectures with the best classification performances (highest NER on the validation set, see next paragraph) have been selected. Since networks are randomly initialised, each network was replicated five times to test its stability. Since no significant differences have been found across the five replicas of the same architecture (figure 5.17), the overall best model considering the NER in validation has been selected. The final parameters for the selected model are shown in table 5.4.

5.2.6 KNN

The kNN algorithm is a benchmark classification method (Altman, 1992). A sample is classified according to the most represented class among the k nearest training samples (neighbors), thus, kNN is also known as a local classification approach. The Euclidean metric has been used for the distance calculation. The optimal number of neighbors k

has been optimized testing different values from 1 to 10, 20, 30, 40 and 50: the optimal k value (4) was found maximizing the NER on the validation set.

5.2.7 Classification measures

The classification performances have been evaluated by means of confusion matrix and derived measures: Non-Error Rate (NER), precision, sensitivity and ROC curves. Details are given in chapter 4

5.2.8 Software

The PARAFAC2 models have been calculated with PARADISe (Johnsen et al., 2017) version 5.8, available at (http://models.life.ku.dk/paradise).

All the classification measures have been calculated by means of routines in the classification toolbox for MATLAB (Ballabio and Consonni, 2013), available at https://michem.unimib.it/download/matlab-toolboxes/classification-toolbox-for-matlab/ (Nov 2, 2021).

The computations, optimization, training and test of the models have been performed in MATLAB (MATLAB 2021a, The MathWorks, Inc. Natick, Massachusetts, United States). The deep neural networks have been calculated with the MATLAB deep network designer toolbox.

5.2.9 Results

The (BI)LSTM, CNN, and kNN classification models were trained with the labeled profiles from the training set and then used to predict profiles included in the external test sets. Neural network hyperparameters and the number of k nearest neighbors were optimized by maximizing the number of correctly classified profiles included in the internal validation set.

The classification performances obtained on the training, validation and test sets are reported in table 5.5. NER values are always higher than 85% for all the models

Table 5.5 – Classification measures on the training, validation and test sets.

Model	Set	NER	Precision Other	Precision Peak	Precision Cutoff peak	Precision Baseline	AUC Other	AUC Peak	AUC Cutoff Peak	AUC Baseline	Sensitivity Other	Sensitivity Peak	Sensitivity Cutoff peak	Sensitivity Baseline
CNN	Train	86.05	86.86	93.20	83.74	89.92	95.30	99.12	99.00	98.19	79.07	94.44	74.19	96.49
	Val	85.53	86.18	92.44	86.23	89.70	95.18	98.89	98.97	98.15	78.94	94.22	72.73	96.21
	Test	88.68	84.16	95.09	97.51	96.39	98.01	96.13	98.34	99.86	96.83	87.32	71.11	99.48
BILSTM	Train	89.40	89.68	90.71	79.52	91.06	95.21	99.31	99.26	98.18	77.10	95.90	88.47	96.13
	Val	89.23	89.16	90.11	81.92	90.75	95.20	99.02	99.24	98.17	76.82	95.30	88.94	95.84
	Test	89.20	88.71	84.33	94.22	91.57	95.81	97.17	99.19	99.91	82.52	89.76	84.65	99.88
KNN	Train	93.84	92.54	91.40	85.51	95.82	98.84	99.90	99.82	99.57	85.86	98.64	95.15	95.72
	Val	90.73	87.98	88.26	83.50	92.90	93.87	98.86	98.55	97.46	78.93	97.31	93.04	93.64
	Test	86.83	92.26	75.17	91.60	88.33	90.74	94.70	94.32	98.34	72.02	93.43	82.96	98.92

when looking at training, validation and test sets. The kNN model has the highest NER considering both training and validation sets, which might be expected, considering the model structure, the low number of neighbors and how the two sets have been produced. However, the kNN model shows the biggest variation in NER values between the training set and the other sets, in particular the NER value decreases 3% and 7% for the validation and the test set, respectively.

The deep learning models are characterized by more stable results between the training and the internal validation set. For the CNN model the variation of the NER for the training set is -0.5% compared to the validation and +2.6% compared to the test set. Considering the BILSTM model, the difference between the training set NER values is -0.2% compared to the internal validation set and -0.2% compared to the test set, respectively. Overall, the performances of the models are satisfactory and the results suggest that all the models achieved excellent classification performances, implying that the classification strategies are appropriate for the task.

In order to further compare the classification performances of the three models on the test set, the aggregated confusion matrix, the NER, the class sensitivities and precisions of the three models achieved on the test set have been considered, as shown in figure 5.18. All the values for these classification measures are listed in table 5.5. In particular, the aggregated confusion matrix derives from the combined analysis of the classification results by the three models. In this plot, the agreement across the three models is reported in a particular representation of a confusion matrix with a Venn-like diagram.
A given profile can be classified as belonging to a given class by i) all the three models (one possible combination), ii) two models (three possible combinations), or iii) just a single model (for a total of three models), for a total of seven possible combinations. Each profile was assigned to one or more of these combinations according to the concordance/discordance of the predictions with respect to the experimental class. In the ixj cell of the aggregated confusion matrix, the number of profiles from i-th class predicted as class j is represented in a graphical way, considering the seven combinations mentioned above. It means that, for each cell, seven numbers, one for

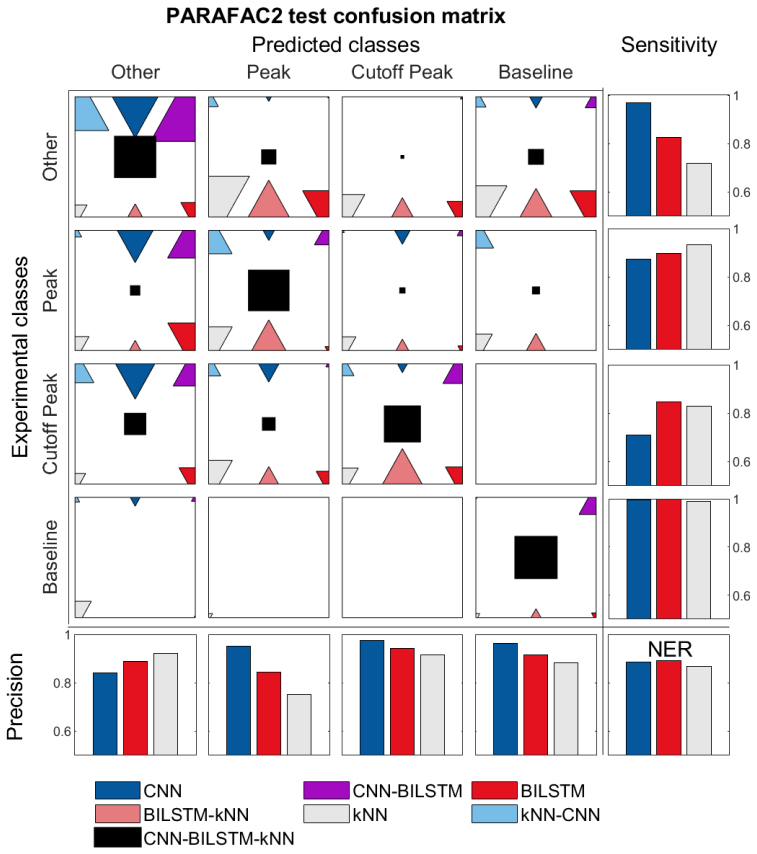

Figure 5.18 – Aggregated confusion matrix for the three approaches (CNN, BILSTM and kNN in blue, red, and white, respectively) calculated for the test set. Details about construction and interpretation are given in the text (Results section). The size of the areas is proportional to the logarithm of the number of profiles. Sensitivities, precisions and NER for each class are also reported as bar plots.

each combination, should be given.

So, for instance, the 1x1 cell of the aggregated confusion matrix reports how many profiles labelled as "Other" were correctly predicted as "Other" by: (1) all the models, i.e., all concordant predictions), (2) CNN and BILSTM models, (3) BILSTM and kNN models, (4) kNN and CNN models, or uniquely by (5) CNN or (6) BILSTM or (7) kNN. In order to ease the visualization, within each cell of the aggregated confusion matrix seven regions have been defined, one for each of the possible combination. In particular, for each cell, the following areas were defined:

1. a black central area that represents the concordant predictions across all the models;

2. a purple area for the CNN and BILSTM models concordant predictions;

3. a pink area for the BILSTM and kNN models concordant predictions;

4. a light blue area for the kNN and CNN models concordant predictions;
5. a blue area for CNN predictions;
6. a red area for BILSTM predictions;
7. a white area for kNN predictions;

The seven areas have been logarithmically rescaled according to the number of the corresponding profiles to help the visual inspection. If all the models give perfect predictions, only the black areas on the main diagonal would be shown in the aggregated confusion matrix. The actual number of profiles for each area is reported in figure 5.19.

From the aggregated confusion matrix represented in figure 5.18, it is possible to see that the black areas have the maximum extent on the diagonal. This suggests that most often the models classify the profiles consistently to each other and that the predictions are correct. This is supported also by the high NER values of the three models reported in the right bottom square (figure 5.18 cell 5x5, table 5.5), where no significant differences can be noticed for the three models.

Considering the sensitivity for the class "Other", the CNN and the BILSTM models perform better compared to the kNN model (cell 1x5, figure 5.18). This can be explained looking at the extent of the pink area in cell 1x1 (figure 5.18) indicating that the CNN and the BILSTM correctly classify many more profiles belonging to this class compared to kNN. It means that the same profiles are assigned to a wrong class by the kNN model. The difference between CNN and BILSTM in terms of sensitivities is due to the set of profiles correctly classified uniquely by the CNN model (blue area in cell 1x1 figure 5.18). Looking at the aggregated confusion matrix, it can be seen that most of the misclassified "Other" profiles by the kNN and BILSTM models are assigned to the class "Peak", as suggested by the greater extent of the white, pink and red areas in the cell 1x2 (figure 5.18). In most of the cases, the profiles resemble a peak but do not fulfill the criteria applied during the labeling. For instance, visually inspecting these profiles show some spikes beyond the 0.1 threshold in addition to the main peak. This could suggest that the CNN is most sensitive to the noise of the profiles, while kNN model is the least sensitive with the BILSTM somewhere in the middle between

the other two methods.

Looking at the precision for the class "Other" obtained by the three models (cell 5x1 figure 5.18), it is possible to notice that the trend is the opposite compared to the sensitivities. In this case the kNN achieved the best result, while the CNN the worst and still the BILSTM in the middle. The smaller precisions for the two deep learning methods are reflected in the aggregated confusion matrix. The great extent of the blue, pink and red areas in the 2x1 and 3x1 cells in figure 5.18 suggests that BILSTM and CNN tend to classify as "Other" more profiles belonging to different classes compared to the kNN model. Visually inspecting these profiles, it has been noticed that in most of the cases there is some residual noise, and the misclassification can be related to that. This observation supports the hypothesis that the deep leaning models are overestimating the impact of noise for the classification of the profiles while the kNN is less sensitive to this aspect. Overall, the BILSTM seems the more balanced model. This different behavior of the models influences the results for the remaining classes, in particular for the "Peak" class. In this case, the kNN has the greatest sensitivity, followed by BILSTM and CNN (cell 2x5, figure 5.18, table 5.5). This is due to the profiles labelled as "Peak" but classified as "Other" by the CNN and BILSTM models (blue, purple and red areas, 2x1 and 3x1 cells, figure 5.18), as discussed before. It means that the kNN model can correctly classify the highest number of profiles labelled as "Peak" compared to BILSTM and CNN. The difference of the sensitivities for the 'Peak' class between the kNN and the CNN is 6.1%. Looking at the precisions for the class "Peak" (cell 5x2, figure 5.18), the trend is the opposite: the CNN has the best precision and the kNN the worst. The white, light blue and pink areas, related to the classifications of the kNN model, are the more extended in the cells 1x2 and 3x2, indicating that kNN tends to assign to the class "Peak" profiles belonging to different classes. In particular, the difference of precisions between the CNN and kNN is 19%. This can be explained considering the criteria adopted during the labelling, where a small difference can discriminate between a class or another. Such small differences between different classes can be problematic to detect for a local model as kNN.

Considering the class "Cutoff peak", the BILSTM model has the highest sensitivity and the CNN the smallest one (cell 3x5, figure 5.18). As for the class "Peak", the

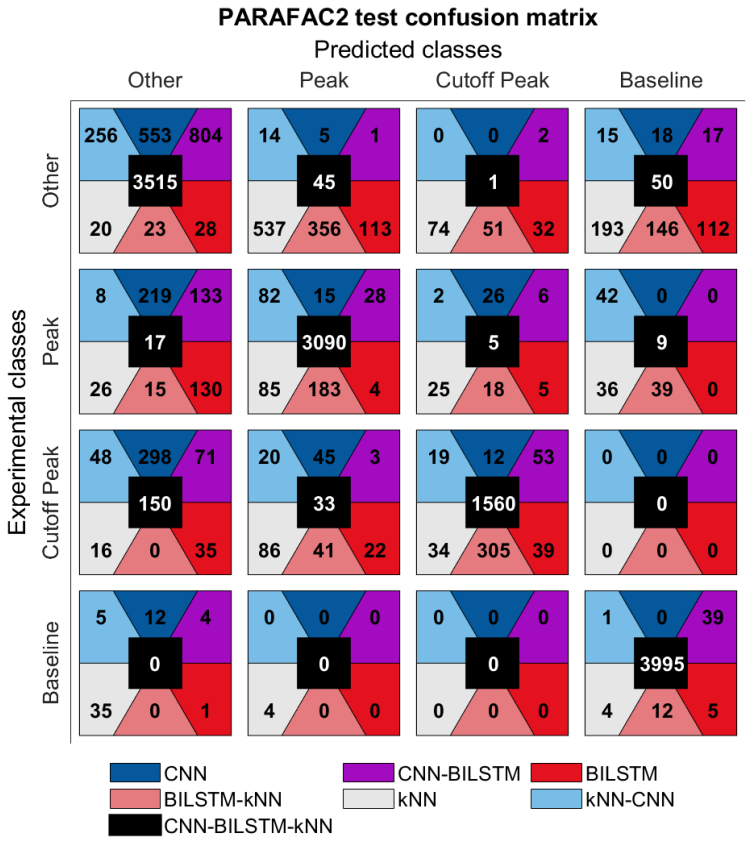

Figure 5.19 – The number of profiles included in the seven areas defined for each cell of the aggregated confusion matrix considering the test set.

sensitivity of the CNN model is lower compared to the other models, because it tends to classify more profiles as "Other". This tendency increases the precision for the CNN model which is the highest also for this class. However, the differences among the three models for this class are less evident and the performances are satisfactory for all the models.

All the three models perform excellently considering the class "Baseline", both in terms of sensitivities and precisions (cells 4x5 and 5x4, respectively, figure 5.18, table 5.5). Looking at the aggregated confusion matrix, the number of profiles classified as "Baseline" from all the three models and actually belonging to this class is 3995 (black area, cell 4x4 figure 5.18 and figure 5.19 A) over a total number of 4056 profiles labelled as "Baseline" in the test set (table 5.3). Thus the 98.4% of the "Baseline" profiles have been correctly classified by all the classification models, moreover almost all the misclassifications (except for only 4 false attributions to peaks by kNN) are due to assignations to the class "Other", i.e. baselines profiles are not confounded with

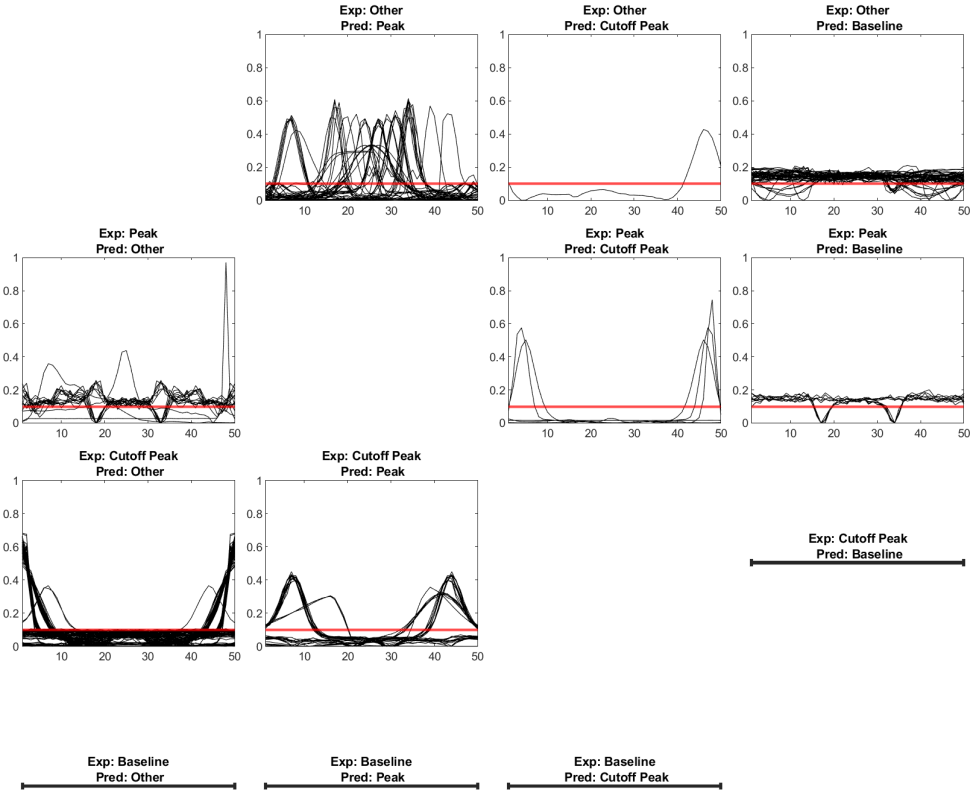

Figure 5.20 – Profiles in the test set misclassified by all the tested models; the red line highlights the 0.1 threshold. The normalized profiles are shown.

peaks or cutoff peaks, further indicating the excellent performances for this class.

A last consideration is about the profiles misclassifed by all the models. Taking advantage of the aggregated confusion matrix, the profiles misclassified by all the models were extracted. These correspond to the profiles represented in the black areas in the off-diagonal cells for a total of 310 profiles (2.02% of all the profiles in the PARAFAC2 test set). All the profiles are shown in figure 5.20. For the cells 3x4 (experimental class: 'Cutoff peak'; predicted: "Baseline") 4x2 (experimental class: "Baseline"; predicted: "Peak") and 4x3 (experimental class: "Baseline"; predicted: "Cutoff peak") no misclassified profiles from all the models were found. In all the other cases, it can be seen that misclassifications mostly depend on borderline profiles (i.e., profiles at the edge between two classes) and by errors in the labelling phase. For instance, in the most representative group corresponding to the cell 3x1 (experimental class: "Cutoff peak"; predicted: "Other") the profiles do not clearly show the inflection point or the tail is slightly over the 0.1 threshold. Extending the same trend also

for the validation and the training sets, some underestimation of the classification performances can be assumed.

5.2.10 Computational time

Table 5.6 – Computational time for the classification of the internal validation set.

Model	Time (seconds)
CNN	4.34
BILSTM	2.43
kNN	64.3

Another result concerns the computational time required by the different models to perform the classification on the profiles of the internal validation set (table 5.6). The calculation has been performed with an Intel® Core™ i7-6950X CPU processor with a dedicated RAM of 32 GigaBytes. The time needed by CNN and the BILSTM models are comparable, the slightly longer running time for the CNN model can be explained by the greater number of hidden layers. The kNN model requires significantly more time compared to the other two models, i.e., it is about 20 times slower. The difference in time required by the models to perform the classification task is important in this context, considering that the analysis of a full GC-MS dataset by means of PARAFAC2 would produce a much higher number of profiles.

5.2.11 ROC curves

In order to further evaluate the classification approaches, the ROC curves for the class 'Peak' from the developed models with the convolutional neural network described in Risum and Bro, 2019 have been compared. The comparison is based on the respective test set. The curves are shown in figure 5.21. Since these results are based on different test sets, the comparison is qualitative with the aim to verify the influence of the data more than the classification performance.

Considering the kNN model, the ROC curve has been calculated considering the posterior probability of a new sample to belong to a given class considering the classes

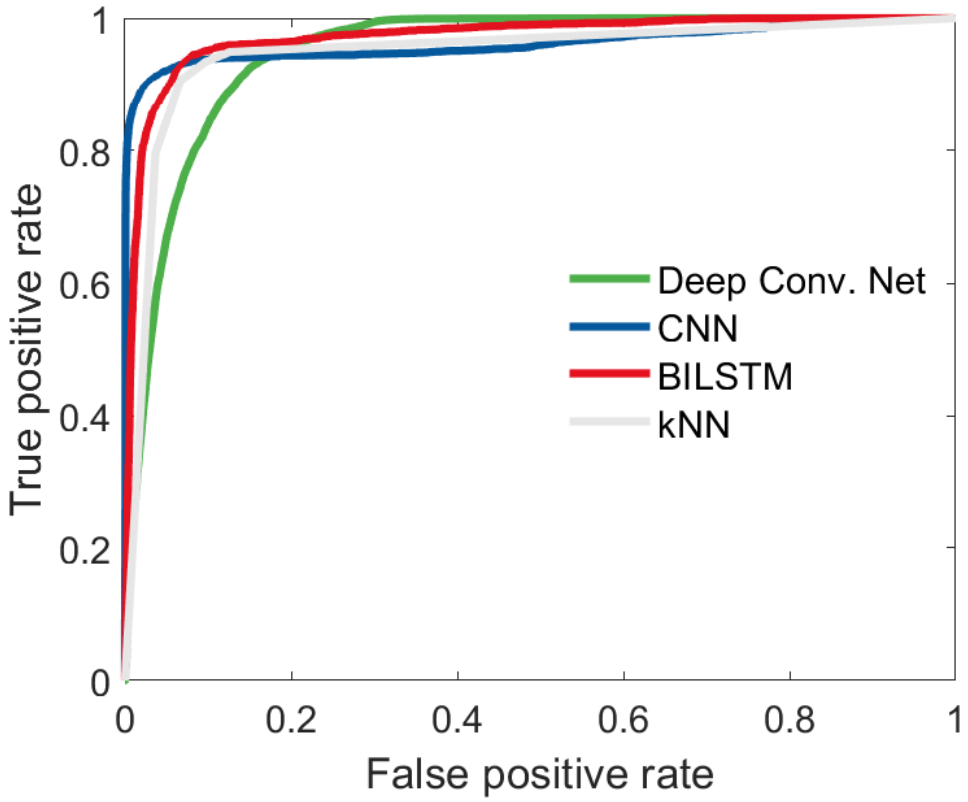

Figure 5.21 – Comparison of the ROC curves for the 'Peak' class of the deep convolutional net from Risum and Bro, 2019 and the proposed CNN, BILSTM and kNN models. The curves are calculated on the respective test sets.

of the k nearest neighbors.

The ROC curve for the deep neural network considering the 'Peak' class in Risum and Bro, 2019 was already next to the top left corner, nonetheless the curves for the three models trained on our data show a further improvement. The AUC reported in Risum and Bro, 2019 was already high, reaching 0.95, however both the CNN and the BILSTM models have an even higher value, 0.96 and 0.97 respectively, suggesting that the adopted criteria for the labelling of the profiles make possible a slight increase of the classification performances. The AUC for the 'Peak' class is 0.95, slightly lower compared to the CNN and the BILSTM models and also considering the previous result. The AUC value for all the classes for the CNN, BILSTM and kNN models is reported in table 5.5.

5.2.12 Conclusions

With the purpose of improving the performances of the CNN proposed in Risum and Bro, 2019, more than 170000 estimated elution profiles have been labelled according to

semi-quantitative criteria, specifically proposed to set a first, preliminary standard for further future developments in this field. The profiles have been used to train again the same CNN architecture and to train two new models based on RNN and KNN. The models have been trained to discriminate among four different classes: 'Peak', 'Cut off peak', 'Baseline' and 'Other'.

The performances of the models have been analyzed by means of an aggregated confusion matrix where the classification results of the three models have been merged. This allowed tracking of the common and non-common classifications, resulting in a deeper insight of the different model trends. Overall, all the three models seem effective. However, there are hints suggesting that the two models based on a deep neural network are 'learning' the underlying criteria applied during the labelling process from the data, while the kNN model seems less robust compared to the other two. Overall, the thorough analysis of the results obtained considering the test set allowed the characterization of the three models, where in general the CNN has the highest precisions and tends to classify more profiles as 'Other', the kNN has the highest sensitivity but fails to properly classify borderline profiles and the BILSTM seems as a compromise between the other two methods. Moreover, the prediction times indicate the two deep learning approach as significantly faster compared to the kNN model. Weighing up all these aspects, the kNN model could be less advisable.

The comparison of the developed models with a previously published deep learning model suggests that it could be possible possible to further improve the classification of elution profiles resolved by PARAFAC2.

Chapter 6

Conclusions and future perspectives

This thesis aims to present and explain a new approach for the identification and quantification of compounds from GC-MS data obtained by untargeted metabolomics experiments.
To introduce the scientific problem of this thesis and the solutions proposed, an overview of the experimental settings, the nature of the resulting data and some of the most common approaches applied by the scientific community to extract the chemical information from GC-MS data have been described.

Nowadays metabolic profiling is an established and widely applied approach. However, the translation of the experimental data into useful information has been indicated several times as the major bottleneck in metabolomics experiments, strongly limiting the full exploitation of the analytical capacity of modern instruments.
In the opinion of the author, this is due to the quite univariate approaches applied in most of the approaches developed for the identification and quantification of the chemical compounds. The data obtained from the analysis of the samples are often analyzed independently of each other. Moreover, each measured mass spectrum is analyzed individually, trying to assess if it describes chemical information.
The identification of the relevant spectra and the estimation of the area under the chromatographic peaks is based on a consolidated theory in the field of chromatography and mass spectrometry. However, the theory was developed considering samples with

much less complexity than those analyzed in metabolic profiling experiments. Considering the complexity of the measured signals, it can be argued that more sophisticated tools are needed to process the data.

Here chemometrics could play a fundamental role. As shown throughout the thesis, the application of multivariate modeling approaches, such as PARAFAC2, can impressively improve the detection of compounds identified from GC-MS data. It has been shown that PARAFAC2 can extract the individual contributions of the different signals resulting in an accurate identification and relative quantification of the chemical compounds detected by the instruments.

However, despite its effectiveness, PARAFAC2 is not a common choice for the analysis of GC-MS data. It is difficult to clearly identify the reasons why, but the need of some knowledge of command line software, such as MATLAB, the language mostly used to implement PARAFAC2, can be recognized among them. Moreover, the modeling results should be visualized and evaluated, and also this can be problematic for less skilled users. Concerning the first point, PARADISe was a major breakthrough to solve this issue, introducing a dedicated GUI but still it requires an expertise in PARAFAC2 modeling to get the final peak table.

The approach proposed in this thesis aims to fill this gap. AutoDise has been designed as an expert system for the automatic handling of the calculation, screening and selection of PARAFAC2 models, in order to automatically extract the components describing the resolved chromatographic peaks, i.e. the pure signals describing the chemical compounds.

A list of components to be included in the peak table is also proposed to the user, selected after a careful screening of the calculated models, applying several diagnostic tools. The screening performed by AutoDise allows the user to focus on a small subset of components, reducing the efforts required in the preliminary phases. A case study was reported to test the efficacy of AutoDise. AutoDise and a traditional software for the analysis of GC-MS data were compared. The results showed that the number of molecules identified by AutoDise is doubled compared to traditional software.

A crucial aspect in this context is the appropriate visualization of the results. Visualization is an essential aspects in chemometrics, and therefore another important

part of this project was the development of a dedicated GUI, in order to provide an intuitive platform for visual analysis of the solutions given by AutoDise. We think it is of primary importance to make the tools developed accessible also to researchers belonging to other research fields.

Overall, the performance of AutoDise is more than satisfactory, but there is still room for improvement. A first improvement was carried out by means of the development of a new ANN model for the identification of the resolved elution profiles. This aspect has been selected given the importance of the CNN developed by (Risum and Bro, 2019) in the AutoDise algorithm. To this end, a thorough manual labeling of more than 170,000 resolved profiles has been performed and new deep learning approaches have been tested showing a further improvement of the classification performances. However, many other developments could be added in the future, in particular for the interval definition step and the selection of the components.

The intervals are defined in order to cover the whole chromatogram. The aim is to extract all the chemical information. However, defining too many intervals leads to a dramatic increase in the time required for the calculation of the models. Thus, a more rational definition of the intervals would help to reduce the calculation time extracting the same information. In this context, one possible way would be to couple the alignment with the interval definition. The alignment phase can be used to identify the relevant regions for the definition of the intervals reducing at the same time the total number of intervals.

Another point that is worth exploring is to integrate other diagnostics for the selection of the components. Among the different possibilities, the more urgent to take into account is the CORCONDIA diagnostic, that has been indicated as an important and reliable parameter for the analysis of PARAFAC2 models, and that is not considered for the screening of the components in AutoDise so far.

It is also important to compare the results of the algorithm described in this thesis with the most applied approaches such as XCMS and MZmine, in order to test the advantages and the weaknesses. I think that in this context an important part would be played by simulated data. To the best of my knowledge, approaches for the simulation of metabolomics GC-MS data has not been proposed. The development of methods

able to simulate the same complexity as in real experimental data would be extremely useful for a more robust comparison of the available software.
Of course, the further developments and improvements of the algorithm and the usability of the GUI could involve many different aspects. However, these tools are expected to be useful to the scientific community. Thus, it would be extremely helpful to get feedback by the users in order to better understand the features to modify or to implement.

A final consideration about this last point is that going through the literature, it can be noticed that the experimental community and the chemometric community, in particular concerning the analysis of GC/LC-MS data, are not communicating much. In many reviews describing the most advanced tools for the identification and/or quantification of chemical compounds for GC-MS data, chemometrics tools such as those described in the present thesis are usually not included. In the opinion of the author, more attention should be placed to better link the experimental groups with the chemometric ones, and this thesis, hopefully, is a little part of this effort.

Appendix I

Appendix

a Description of autodise-table.xlsx file

Details about the .xlsx file generated by the AutoDise GUI are given in this section.

First sheet

In the first sheet, called "Overview" general information is reported :

Paradise version (currently 6.0.0)

Export Date The day the file has been generated

Number of samples

Mass spec. range Range of the m/z in the raw data.

Retention time range (min)

Preprocessing Indicates whether the data have been aligned

Interval Selection Manual or auto, depending on the user selection

Modelling Model applied to resolve the data

Chemical Identification Description of the second sheet contents

Relative Concentrations Description of the third sheet contents

Top NIST hits Description of the forth sheet contents

Resolved Mass Spectra Description of the fifth sheet contents

Interval Details Description of the sixth sheet contents

Second sheet

In the second sheet, called "Relative concentrations", the estimated relative concentration are reported. It is organized as a table:

First row Headers of the columns in the table

- **First column "Compound name"** Name of the identified molecule.
- **Second column "Match quality"** Quality of the match of the resolved spectra with the NIST library.
- **Third column "Compound ID"** ID of the corresponding compound.
- **Fourth column "Interval ID"** Interval (indicated with the number) from where the compound was resolved.
- **Fifth column "Est. Retention Time (min)"** Retention time of the compound
- **Sixth column - Last column "Samples #"** From the sixth to the last column the relative concentration among the samples is reported. Each column corresponds to a sample.

Second row - Last row Each row corresponds to one of the compounds identified from AutoDise

Third sheet

In the third sheet, called "Resolved Mass Spectra", the estimated mass spectra of each identified compound are reported. It is organized as a table:

First row Headers of the columns in the table

- **First column "Compound name"** Name of the identified molecule.
- **Second column "Match quality"** Quality of the match of the resolved spectra with the NIST library.

Third column "Compound ID" ID of the corresponding compound.

Fourth column "Interval ID" Interval (indicated with the number) from where the compound was resolved.

Fifth column - Last column "mz #" From the fifth to the last column the estimated values for each mz are reported. Each column corresponds to a specific m/z.

Second row - Last row Each row corresponds to one of the compounds identified from AutoDise

Fourth sheet

In the fourth sheet, called "Top NIST hits", the two best hits for each identified compound are reported. It is organized as a table:

First row Headers of the columns in the table

First column "Compound name" Name of the identified molecule.

Second column "Match quality" Quality of the match of the resolved spectra with the NIST library.

Third column "Compound ID" ID of the corresponding compound.

Fourth column "Interval ID" Interval (indicated with the number) from where the compound was resolved.

Fifth column "Est. Retention Time (min)" Retention time of the compound

Sixth column "Hit 1: Compound" Putative compound with the best hit

Seventh column "Hit 1: Structure" Molecular structure of the putative compound

Eighth column "Hit 1: Match Factor"

Ninth column "CAS" CAS number of the putative compound

Tenth column "Hit 1: Mw" Molecular weight of the putative compound

Eleventh column - last column Equal to Fifth column - Tenth column but for the second best hit.

Second row - Last row Each row corresponds to one of the compounds identified from AutoDise

Fifth sheet

In the fifth sheet, called "Intervals details", diagnostics about the models where compound(s) have been extracted are reported. It is organized as a table:

First row Headers of the columns in the table

First column "Interval ID" Interval (indicated with the number) from where the compound was resolved

Second column "Model Components (max. size)" Number of components included in the model

Third column "Chemical Components" Number of components extracted by the model

Fourth column "Num. Iterations" Number of iterations to fit the model

Fifth column "Compute Time (sec.)" Time to fit the model

Sixth column "Core Consistency"

Seventh column "SSTotal" Residuals value

Eighth column "VarExpl" Explained variance by the model

Second row - Last row Each row corresponds to one of the compounds identified from AutoDise

Bibliography

Abdollahi, Hamid and Romà Tauler (2011). "Uniqueness and rotation ambiguities in multivariate curve resolution methods". In: *Chemometrics and Intelligent Laboratory Systems* 108.2, pp. 100–111.

Alseekh, Saleh and Alisdair R Fernie (2018). "Metabolomics 20 years on: what have we learned and what hurdles remain?" In: *The Plant Journal* 94.6, pp. 933–942.

Alseekh, Saleh et al. (2021). "Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices". In: *Nature methods* 18.7, pp. 747–756.

Altman, Naomi S (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". In: *The American Statistician* 46.3, pp. 175–185.

Amigo, José Manuel, Thomas Skov, and Rasmus Bro (2010). "ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics". In: *Chemical reviews* 110.8, pp. 4582–4605.

Anderson, Paul E et al. (2008). "Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics". In: *Metabolomics* 4.3, pp. 261–272.

Anderson, Paul E et al. (2011). "Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data". In: *Metabolomics* 7.2, pp. 179–190.

Araújo, Ana Margarida et al. (2018). "GC–MS metabolomics reveals disturbed metabolic pathways in primary mouse hepatocytes exposed to subtoxic levels of 3, 4-methylenedioxymethamphetamine (MDMA)". In: *Archives of toxicology* 92.11, pp. 3307–3323.

Arbona, Vicent et al. (2009). "Plant phenotype demarcation using nontargeted LC-MS and GC-MS metabolite profiling". In: *Journal of agricultural and food chemistry* 57.16, pp. 7338–7347.

Ausloos, P et al. (1999). "The critical evaluation of a comprehensive mass spectral library". In: *Journal of the American Society for Mass Spectrometry* 10.4, pp. 287–299.

Baccolo, Giacomo et al. (2021). "From untargeted chemical profiling to peak tables–A fully automated AI driven approach to untargeted GC-MS". In: *TrAC Trends in Analytical Chemistry* 145, p. 116451.

Ballabio, Davide and Viviana Consonni (2013). "Classification tools in chemistry. Part 1: linear models. PLS-DA". In: *Analytical Methods* 5.16, pp. 3790–3798.

Ballabio, Davide, Francesca Grisoni, and Roberto Todeschini (2018). "Multivariate comparison of classification performance measures". In: *Chemometrics and Intelligent Laboratory Systems* 174, pp. 33–44.

Ballabio, Davide et al. (2008). "Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 22.8, pp. 457–463.

Beale, David J et al. (2018). "Review of recent developments in GC–MS approaches to metabolomics-based research". In: *Metabolomics* 14.11, pp. 1–31.

Berg, Robert A van den et al. (2006). "Centering, scaling, and transformations: improving the biological information content of metabolomics data". In: *BMC genomics* 7.1, pp. 1–15.

Berge, Jos MF ten and Henk AL Kiers (1996). "Some uniqueness results for PARAFAC2". In: *Psychometrika* 61.1, pp. 123–132.

Bicking, Merlin KL (2006). "Integration errors in chromatographic analysis, Part I: peaks of approximately equal size". In: *LC-GC North America* 24.4, pp. 402–410.

Booksh, Karl S and Bruce R Kowalski (1994). "Theory of analytical chemistry". In: *Analytical Chemistry* 66.15, 782A–791A.

Brack, Werner et al. (2016). "Effect-directed analysis supporting monitoring of aquatic environments—an in-depth overview". In: *Science of the Total Environment* 544, pp. 1073–1118.

Bro, Rasmus (1998). "Multi-way analysis in the food industry-models, algorithms, and applications". In: *MRI, EPG and EMA," Proc ICSLP 2000*. Citeseer.

Bro, Rasmus and Henk AL Kiers (2003). "A new efficient method for determining the number of components in PARAFAC models". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 17.5, pp. 274–286.

Bundy, Jacob G, Matthew P Davey, and Mark R Viant (2009). "Environmental metabolomics: a critical review and future perspectives". In: *Metabolomics* 5.1, pp. 3–21.

Cattell, Raymond B and AKS Cattell (1955). "Factor rotation for proportional profiles: analytical solution and an example". In: *British Journal of Statistical Psychology* 8.2, pp. 83–92.

Cavanna, Daniele et al. (2018). "The scientific challenges in moving from targeted to non-targeted mass spectrometric methods for food fraud analysis: A proposed validation workflow to bring about a harmonized approach". In: *Trends in Food Science & Technology* 80, pp. 223–241.

Chaleckis, Romanas et al. (2019). "Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics". In: *Current opinion in biotechnology* 55, pp. 44–50.

Clarke, Robert et al. (2008). "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data". In: *Nature reviews cancer* 8.1, pp. 37–49.

Coble, Jamie B and Carlos G Fraga (2014). "Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery". In: *Journal of chromatography A* 1358, pp. 155–164.

Cohen, Jeremy E and Rasmus Bro (2018). "Nonnegative PARAFAC2: A flexible coupling approach". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 89–98.

Cox, Michael M and David L Nelson (2008). *Lehninger principles of biochemistry*. Vol. 5. Wh Freeman New York.

Cubero-Leon, Elena, Rosa Peñalver, and Alain Maquet (2014). "Review on metabolomics for food authentication". In: *Food Research International* 60, pp. 95–107.

D'Arcy, Peter and W Gary Mallard (2004). "AMDIS–user guide". In: *US Department of Commerce, Technology Administration, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.[Google Scholar]*.

Davis, Richard A et al. (2007). "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform". In: *Chemometrics and intelligent laboratory systems* 85.1, pp. 144–154.

De Juan, Anna, Joaquim Jaumot, and Romà Tauler (2014). "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem". In: *Analytical Methods* 6.14, pp. 4964–4976.

De Juan, Anna and Romà Tauler (2006). "Multivariate curve resolution (MCR) from 2000: progress in concepts and applications". In: *Critical Reviews in Analytical Chemistry* 36.3-4, pp. 163–176.

De Leeuw, Jan, Forrest W Young, and Yoshio Takane (1976). "Additive structure in qualitative data: An alternating least squares method with optimal scaling features". In: *Psychometrika* 41.4, pp. 471–503.

De Meyer, Tim et al. (2008). "NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm". In: *Analytical chemistry* 80.10, pp. 3783–3790.

DeJong, Stephanie et al. (2016). *Correcting Saturation in Mass Spectrometry Data using Principal Components Analysis.* Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Diederen, Tomek et al. (2021). "Metabolomics". In: *Metabolic Engineering: Concepts and Applications* 13, pp. 259–299.

Domingo-Almenara, Xavier et al. (2016). "eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS-based metabolomics". In: *Analytical Chemistry* 88.19, pp. 9821–9829.

Dromey, RG et al. (1976). "Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data". In: *Analytical Chemistry* 48.9, pp. 1368–1375.

Dunn, Warwick B et al. (2013). "Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics". In: *Metabolomics* 9.1, pp. 44–66.

Dyson, Norman Allen and Roger M Smith (1998). *Chromatographic integration methods.* Vol. 3. Royal Society of Chemistry.

Emwas, Abdul-Hamid M (2015). "The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research". In: *Metabonomics*. Springer, pp. 161–193.

Engelsen, Søren Balling, PS Belton, and Hans Jørgen Jakobsen (2005). *Magnetic resonance in food science: the multivariate challenge*. Vol. 299. Royal Society of Chemistry.

European Chemicals Agency (2016). *New approach methodologies in regulatory science - Publications Office of the EU*. Tech. rep. URL: https://op.europa.eu/en/publication-detail/-/publication/4c2ad7eb-9ad0-11e6-868c-01aa75ed71a1/language-en.

European Food Safety (2014). "Modern methodologies and tools for human hazard assessment of chemicals". In: *EFSA Journal* 12.4, p. 3638.

Garreta-Lara, Elba et al. (2016). "Metabolic profiling of Daphnia magna exposed to environmental stressors by GC–MS and chemometric tools". In: *Metabolomics* 12.5, p. 86.

Giordani, Paolo, Roberto Rocci, and Giuseppe Bove (2020). "Factor Uniqueness of the Structural Parafac Model". In: *psychometrika* 85.3, pp. 555–574.

Golub, Gene and William Kahan (1965). "Calculating the singular values and pseudo-inverse of a matrix". In: *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 2.2, pp. 205–224.

Goodacre, Royston et al. (2007). "Proposed minimum reporting standards for data analysis in metabolomics". In: *Metabolomics* 3.3, pp. 231–241.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.

HARSHMAN, RA (1970). "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-mode factor analysis". In: *UCLA Working Papers in Phonetics* 16, pp. 1–84.

Harshman, RA and ME Lundy (1984). *Research methods in multimode data analysis, chapter The PARAFAC model for three-way factor analysis and multidimensional scaling*.

Harshman, Richard A et al. (1972). "PARAFAC2: Mathematical and technical notes". In: *UCLA working papers in phonetics* 22.3044, p. 122215.

Hirschfeld, Tomas (1980). "The Hy-phen-ated Methods". In: *Analytical Chemistry* 52.2, 297A–312A.

Hochreiter, Sepp (1998). "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02, pp. 107–116.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Huan, Tao et al. (2017). "Systems biology guided by XCMS Online metabolomics". In: *Nature methods* 14.5, pp. 461–462.

Huang, Shao-Min et al. (2013). "Metabolomics of developing zebrafish embryos using gas chromatography-and liquid chromatography-mass spectrometry". In: *Molecular BioSystems* 9.6, pp. 1372–1380.

Jacob, Minnie et al. (2019). "Metabolomics toward personalized medicine". In: *Mass spectrometry reviews* 38.3, pp. 221–238.

Johnsen, Lea G et al. (2014). "Automated resolution of overlapping peaks in chromatographic data". In: *Journal of Chemometrics* 28.2, pp. 71–82.

Johnsen, Lea G et al. (2017). "Gas chromatography–mass spectrometry data processing made easy". In: *Journal of Chromatography A* 1503, pp. 57–64.

Johnson, Caroline H, Julijana Ivanisevic, and Gary Siuzdak (2016). "Metabolomics: beyond biomarkers and towards mechanisms". In: *Nature reviews Molecular cell biology* 17.7, pp. 451–459.

Kamstrup-Nielsen, Maja H, Lea G Johnsen, and Rasmus Bro (2013). "Core consistency diagnostic in PARAFAC2". In: *Journal of Chemometrics* 27.5, pp. 99–105.

Kapoore, Rahul Vijay and Seetharaman Vaidyanathan (2016). "Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2079, p. 20150363.

Katajamaa, Mikko and Matej Orešič (2005). "Processing methods for differential analysis of LC/MS profile data". In: *BMC bioinformatics* 6.1, pp. 1–12.

Kiers, Henk AL and Age K Smilde (1995). "Some theoretical results on second-order calibration methods for data with and without rank overlap". In: *Journal of chemometrics* 9.3, pp. 179–195.

Koek, Maud M et al. (2011). "Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives". In: *Metabolomics* 7.3, pp. 307–328.

Koh, Yueting et al. (2010). "Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data". In: *Journal of chromatography A* 1217.52, pp. 8308–8316.

Kowalski, B R_ and CF Bender (1972). "k-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation". In: *Analytical Chemistry* 44.8, pp. 1405–1411.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25, pp. 1097–1105.

Kruskal, JB (1983). "Multilinear methods". In: *Proc. Symp. Appl. Math.* Vol. 28, p. 75.

Kuhl, Carsten et al. (2012). "CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets". In: *Analytical chemistry* 84.1, pp. 283–289.

Lai, Zijuan et al. (2018). "Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics". In: *Nature methods* 15.1, pp. 53–56.

Lassen, Johan et al. (2021). "Assessment of XCMS Optimization Methods with Machine-Learning Performance". In: *Analytical Chemistry* 93.40, pp. 13459–13466.

Laursen, Martin F et al. (2020). "Breastmilk-promoted bifidobacteria produce aromatic lactic acids in the infant gut". In: *BioRxiv*.

Lawton, William H and Edward A Sylvestre (1971). "Self modeling curve resolution". In: *Technometrics* 13.3, pp. 617–633.

Li, Yinbo and Gonzalo R Arce (2004). "A maximum likelihood approach to least absolute deviation regression". In: *EURASIP Journal on Advances in Signal Processing* 2004.12, pp. 1–8.

Li, Zhucui et al. (2018). "Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection". In: *Analytica chimica acta* 1029, pp. 50–57.

Libiseller, Gunnar et al. (2015). "IPO: a tool for automated optimization of XCMS parameters". In: *BMC bioinformatics* 16.1, pp. 1–10.

Lisec, Jan et al. (2006). "Gas chromatography mass spectrometry–based metabolite profiling in plants". In: *Nature protocols* 1.1, pp. 387–396.

Liu, Huijun et al. (2014). "Characterization of volatile organic metabolites in lung cancer pleural effusions by SPME–GC/MS combined with an untargeted metabolomic method". In: *Chromatographia* 77.19-20, pp. 1379–1386.

Lommen, Arjen (2009). "MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing". In: *Analytical chemistry* 81.8, pp. 3079–3086.

Lorenzo-Seva, Urbano and Jos MF Ten Berge (2006). "Tucker's congruence coefficient as a meaningful index of factor similarity". In: *Methodology* 2.2, pp. 57–64.

Malmquist, Gunnar and Rolf Danielsson (1994). "Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods". In: *Journal of Chromatography A* 687.1, pp. 71–88.

Markley, John L et al. (2017). "The future of NMR-based metabolomics". In: *Current opinion in biotechnology* 43, pp. 34–40.

May, Jody C and John A McLean (2016). "Advanced multidimensional separations in mass spectrometry: navigating the big data deluge". In: *Annual review of analytical chemistry* 9, pp. 387–409.

Misra, Biswapriya B (2018). "New tools and resources in metabolomics: 2016–2017". In: *Electrophoresis* 39.7, pp. 909–923.

— (2021). "New software tools, databases, and resources in metabolomics: updates from 2020". In: *Metabolomics* 17.5, pp. 1–24.

Myers, Owen D et al. (2017). "Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data". In: *Analytical Chemistry* 89.17, pp. 8689–8695.

NIST (2020). *NIST20: Updates to the NIST Tandem and Electron Ionization Spectral Libraries*. URL: https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries.

Niu, Weihuan et al. (2014). "Comparative evaluation of eight software programs for alignment of gas chromatography–mass spectrometry chromatograms in metabolomics experiments". In: *Journal of Chromatography A* 1374, pp. 199–206.

Olivieri, Alejandro C (2021). "A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution- a tutorial". In: *Analytica Chimica Acta*, p. 338206.

O'Shea, Keiron and Biswapriya B Misra (2020). "Software tools, databases and resources in metabolomics: updates from 2018 to 2019". In: *Metabolomics* 16.3, pp. 1–23.

Pinto, Joana et al. (2018). "Assessment of oxidation compounds in oaked Chardonnay wines: A GC–MS and 1H NMR metabolomics approach". In: *Food chemistry* 257, pp. 120–127.

Pluskal, Tomáš et al. (2010). "MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data". In: *BMC bioinformatics* 11.1, pp. 1–11.

Quintanilla-Casas, Beatriz et al. (2020). "Supporting the sensory panel to grade virgin olive oils: an In-House-Validated screening tool by volatile fingerprinting and chemometrics". In: *Foods* 9.10, p. 1509.

Rajkó, Róbert et al. (2017). "On uniqueness of the non-negative decomposition of two-and three-component three-way data arrays". In: *Chemometrics and Intelligent Laboratory Systems* 160, pp. 91–98.

Rattray, Nicholas JW et al. (2014). "Taking your breath away: metabolomics breathes life in to personalized medicine". In: *Trends in biotechnology* 32.10, pp. 538–548.

Risum, Anne Bech and Rasmus Bro (2019). "Using deep learning to evaluate peaks in chromatographic data". In: *Talanta* 204, pp. 255–260.

Roberts, Lee D et al. (2012). "Targeted metabolomics". In: *Current protocols in molecular biology* 98.1, pp. 30–2.

Rosato, Antonio et al. (2018). "From correlation to causation: analysis of metabolomics data using systems biology approaches". In: *Metabolomics* 14.4, pp. 1–20.

Røst, Lisa M et al. (2020). "Absolute quantification of the central carbon metabolome in eight commonly applied prokaryotic and eukaryotic model systems". In: *Metabolites* 10.2, p. 74.

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747*.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Sales, C et al. (2019). "Olive oil quality classification and measurement of its organoleptic attributes by untargeted GC–MS and multivariate statistical-based approach". In: *Food chemistry* 271, pp. 488–496.

Savitzky, Abraham and Marcel JE Golay (1964). "Smoothing and differentiation of data by simplified least squares procedures." In: *Analytical chemistry* 36.8, pp. 1627–1639.

Schrimpe-Rutledge, Alexandra C et al. (2016). "Untargeted metabolomics strategies—challenges and emerging directions". In: *Journal of the American Society for Mass Spectrometry* 27.12, pp. 1897–1905.

Sévin, Daniel C et al. (2015). "Biological insights through nontargeted metabolomics". In: *Current opinion in biotechnology* 34, pp. 1–8.

Sharma, Sagar, Simone Sharma, and Anidhya Athaiya (2017). "Activation functions in neural networks". In: *towards data science* 6.12, pp. 310–316.

Skoog, Douglas A et al. (2013). *Fundamentals of analytical chemistry*. Cengage learning.

Skov, Thomas et al. (2006). "Automated alignment of chromatographic data". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 20.11-12, pp. 484–497.

Smilde, Age, Rasmus Bro, and Paul Geladi (2005). *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons.

Smirnov, Aleksandr et al. (2018). "ADAP-GC 3.2: graphical software tool for efficient spectral deconvolution of gas chromatography–high-resolution mass spectrometry metabolomics data". In: *Journal of proteome research* 17.1, pp. 470–478.

Smith, Colin A et al. (2006). "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification". In: *Analytical chemistry* 78.3, pp. 779–787.

Sparkman, O David, Zelda Penton, and Fulton G Kitson (2011). *Gas chromatography and mass spectrometry: a practical guide*. Academic press.

Spicer, Rachel et al. (2017). "Navigating freely-available software tools for metabolomics analysis". In: *Metabolomics* 13.9, pp. 1–16.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Stein, Stephen E (1999). "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data". In: *Journal of the American Society for Mass Spectrometry* 10.8, pp. 770–781.

Stein, Stephen E and Donald R Scott (1994). "Optimization and testing of mass spectral library search algorithms for compound identification". In: *Journal of the American Society for Mass Spectrometry* 5.9, pp. 859–866.

Stoyanova, R and TR Brown (2001). "NMR spectral quantitation by principal component analysis". In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 14.4, pp. 271–277.

Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney (2012). "LSTM neural networks for language modeling". In: *Thirteenth annual conference of the international speech communication association.*

Tauler, Roma (1995). "Multivariate curve resolution applied to second order data". In: *Chemometrics and intelligent laboratory systems* 30.1, pp. 133–146.

Tauler, Roma and Hadi Parastar (2018). "Big (Bio) chemical data mining using chemometric methods: a need for chemists". In: *Angewandte Chemie International Edition.*

Tautenhahn, Ralf et al. (2012). "XCMS Online: a web-based platform to process untargeted metabolomic data". In: *Analytical chemistry* 84.11, pp. 5035–5039.

Ten Berge, Jos MF and Nikolaos D Sidiropoulos (2002). "On uniqueness in CANDECOMP/PARAFAC". In: *Psychometrika* 67.3, pp. 399–409.

Tomasi, Giorgio, Frans Van Den Berg, and Claus Andersson (2004). "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 18.5, pp. 231–241.

Tsugawa, Hiroshi et al. (2015). "MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis". In: *Nature methods* 12.6, pp. 523–526.

Tucker, Ledyard R (1951). *A method for synthesis of factor analysis studies.* Tech. rep. Educational Testing Service Princeton Nj.

Ueberschaar, Nico et al. (2021). "Soil Solution Analysis With Untargeted GC–MS—A Case Study With Different Lysimeter Types". In: *Frontiers in Earth Science* 8, p. 686.

Want, Elizabeth and Perrine Masson (2011). "Processing and analysis of GC/LC-MS-based metabolomics data". In: *Metabolic profiling*. Springer, pp. 277–298.

Wehrens, Ron, Georg Weingart, and Fulvio Mattivi (2014). "metaMS: an open-source pipeline for GC–MS-based untargeted metabolomics". In: *Journal of Chromatography B* 966, pp. 109–116.

Williams, RJ and HK Berry (1951). "Individual metabolic patterns and human disease: an exploratory study utilizing predominantly paper chromatographic methods". In: *Introduction, General Discussion and Tentative Conclusions*, pp. 7–21.

Wilson, ID and UA Th Brinkman (2003). "Hyphenation and hypernation: the practice and prospects of multiple hyphenation". In: *Journal of Chromatography A* 1000.1-2, pp. 325–356.

Wishart, David S (2008). "Quantitative metabolomics using NMR". In: *TrAC trends in analytical chemistry* 27.3, pp. 228–237.

— (2016). "Emerging applications of metabolomics in drug discovery and precision medicine". In: *Nature reviews Drug discovery* 15.7, pp. 473–484.

— (2019). "NMR metabolomics: A look ahead". In: *Journal of Magnetic Resonance* 306, pp. 155–161.

Wold, Herman (1975). "Path models with latent variables: The NIPALS approach". In: *Quantitative sociology*. Elsevier, pp. 307–357.

Wülfert, Florian, Wim Th Kok, and Age K Smilde (1998). "Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models". In: *Analytical chemistry* 70.9, pp. 1761–1767.

Young, Forrest W, Yoshio Takane, and Jan de Leeuw (1978). "The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features". In: *Psychometrika* 43.2, pp. 279–281.

Yu, Huiwen and Rasmus Bro (2021). "PARAFAC2 and local minima". In: *Chemometrics and Intelligent Laboratory Systems* 219, p. 104446.

Yuille, Samantha et al. (2018). "Human gut bacteria as potent class I histone deacetylase inhibitors in vitro through production of butyric acid and valeric acid". In: *PloS one* 13.7, e0201073.

Zhou, Bin et al. (2012). "LC-MS-based metabolomics". In: *Molecular BioSystems* 8.2, pp. 470–481.