THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université

le 17 décembre 2021 par

Elena Piersanti

Application and development of NMR methods and signal processing for the analysis of small molecule mixtures

Discipline
Sciences chimiques

Discipline
Chimie analytique

École doctorale
ED 250 - Sciences chimiques de Marseille

Laboratoire/Partenaires de recherche
Aix Marseille Université
Institut des Sciences Moléculaires de Marseille (iSm2)
Institut des Mathématiques de Marseille (I2M)

Composition du jury

- Dr. Jonathan FARJON   Rapporteur
Université de Nantes

- Pr. Sabine BOUGUET-BONNET   Rapporteuse
Université de Lorraine

- Pr. Jean-Marc NUZILLARD   Examinateur
Université de Reims Champagne-Ardenne

- Pr. Bruno TORRESANI   Président du jury
Université d'Aix-Marseille

- Dr. Laurent DUVAL   Examinateur
IFP - Energies Nouvelles

- Pr. Mylène CAMPREDON   Directrice de thèse
Université d'Aix-Marseille

- Dr. Mehdi YEMLOUL   Co-encadrant de thèse
Université d'Aix-Marseille

Affidavit

I, undersigned, Elena PIERSANTI, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific direction of Mylène CAMPREDON, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place : Marseille, date : October 25, 2021

Table of contents

List of Notations and Abbreviations

Δ: diffusion delay

δ: gradient pulse duration

$\vec{I}$: Spin moment

γ: Gyromagnetic ratio

λ: thresholding value

$\vec{\mu}$: magnetic moment

1D, 2D, nD: One, two, multi-dimensional NMR

$\overrightarrow{B_0}$: Magnetic field

BC-VMFB: Block-Coordinate Variable Metric Forward–Backward

BSS: Blind Source Separation

COSY: Correlation SpectroscopY

CPMG: Carr Purcell Meiboom Gill

D: Diffusion coefficient (m^2/s)

DOSY: Diffusion Ordered SpectroscopY

DW: Dwell time

FID: Free Induction Decay

FT: Fourier Transform

G: Gradient Strength

HSQC: Heteronuclear Single Quantum Coherence

HR-MAS: High Resolution – Magic Angle Spinning

ILT: Inverse Laplace Transformation

INEPT: Insensitive Nuclei Enhanced by Polarization Transfer

J: Scalar couping constant

JADE: Joint Approximate Diagonalization of Eigen-matrices

k: number of sources

LC: Liquid Chromatography

LIM: Linear Instantaneous Mixture

m: number of mixtures

$\vec{M}$: Nuclear magnetization vector

MAS: Magic Angle Spinning

MaxQ: Maximum Quantum

MQ: Multiple Quantum

MS: Mass Spectrometry

Nc_proc: Intensity scaling factor

NMF: Non negative Matrix Factorization

NMR: Nuclear Magnetic Resonance

NS: Number of Scans

PALM: Proximal alternating linearized minimization

PALS: Projected Alternate Least Squares

PCA: Principal Component Analysis

PGSE: Pulsed Gradient Spin-Echo

PROJECT: Periodic Refocusing of J Evolution by Coherence Transfer

RF: Radio-Frequency

SDR: Signal to Distortion Ratio

SIR: Signal to Interference Ratio

SNR: Signal-to-Noise Ratio

STALS: Soft Thresholded Projected Alternate Least Squares

SW: Spectral Width

T_1: Longitudinal relaxation time

T_2: Transverse relaxation time

$T_{1\rho}$: Spin lattice relaxation in the rotating frame

zg: 1D Bruker pulse program

BIFROST Project and Funding Support

The work presented in this thesis is part of the BIFROST project funded by the Excellence Initiative of Aix-Marseille University – A*Midex foundation, a French "Investissements d'Avenir" program. It involves three partners: Institut des Sciences Moléculaires de Marseille (ISM2) (with chemists), Institut des Mathematiques de Marseille (I2M) (for mathematical part) and an industrial partner, IFP-Energies Nouvelles.

BIFROST (Blind Identification, Filtering & Restoration On Spectral Techniques) concerns the development of data processing and data acquisition schemes able to improve the power of chemical analysis of complex mixtures, in view of achieving a qualitative and quantitative decomposition of instrumental responses, which are the NMR spectra by addressing bottlenecks in both disciplines. The challenge consists in generating stable algorithms producing high-purity source representation in presence of signal distortions and instabilities and, more importantly, a wide dynamic range of molecular concentrations. This last aspect is crucial for several reasons. From the chemical point of view, the most abundant compounds do not always carry relevant information about the state of a sample, with biomarkers and contaminants being typical examples of this. The most intricate case of study is the one in which the spectrum presents severe overlaps, so when more intense signals obscure those ones from minor species. Consequently, an obvious possibility of investigation, improving the detection of less abundant compounds, is to try to find analytical techniques with increased resolution. Multidimensional (nD) analyses are a possible response to this approach that has been extensively researched. Interestingly, signal processing, namely covariance analysis, has been used with some success to create the equivalent of nD spectra by relying on variations of the signal intensity along series of samples. From the mathematical viewpoint, the unmixing problem is a blind source separation problem, which can also be seen as an instance of the dictionary learning problem, which currently receives considerable attention. Dictionary learning generally leads to difficult non-convex optimization problems, for which there exist very few probably convergent and stable algorithms. A main goal of this project was to develop mathematical and signal processing approaches which stay as close as possible to signal acquisition, *e.g.* avoiding pre-processing methods and software.

OUTLINE

Nuclear Magnetic Resonance spectroscopy is a powerful tool to analyze complex mixtures where it is possible to have structural and quantitative information within one spectrum. Despite the development of NMR methods, the growing complexity of the samples leads to crowded spectra that compromise the analytical performances of this technique. One-dimensional NMR spectra can become unusable and the interpretation of two-dimensional spectra is often very difficult and time-consuming. The association of new mathematical methods for signal processing with the methodological developments in NMR is a promising evolution. In this context, the first part of my thesis is focused on the analysis of small molecule mixtures by NMR with the application of Blind Source Separation (BSS)[1] algorithms. This source separation technique, originally used for other disciplines such as telecommunications and neurosciences,[2,3] has shown its effectiveness for the de-mixing of 1D and 2D NMR spectra.[4] Spectral decomposition is performed using correlations, essentially variation in concentrations detected over a series of data sets, which allows the extraction of the spectra of compounds in mixtures.[5] [6] Several types of samples, synthetic mixtures of terpenes and many others, are used to evaluate the efficiency of the algorithms.

The second part of my thesis is concerned with two additional developments which can be defined as side projects: (i) the implementation and assessment of $T_{1\rho}$ relaxation filter in metabolomics to remove the broad signals of macromolecules which are usually not taken into account; (ii) the application of the well-established 2D Maximum Quantum (MaxQ NMR) NMR approach to simplify the NMR spectra of mixtures resulting from the reaction of enzyme complexes.

Keywords: NMR spectroscopy, signal processing, mixture analysis

APERÇU GÉNÉRAL DE LA THÈSE

La spectroscopie par résonance magnétique nucléaire (RMN) est un outil puissant pour analyser des mélanges complexes où il est possible d'avoir des informations structurales et quantitatives à partir d'un seul spectre. Malgré le développement des méthodes RMN, la complexité croissante des échantillons conduit à des spectres encombrés qui compromettent les performances analytiques de cette technique. Les spectres RMN monodimensionnels peuvent devenir inexploitables et l'interprétation des spectres bidimensionnels est souvent très difficile et chronophage. L'utilisation de nouvelles méthodes mathématiques de traitement du signal associées aux développements méthodologiques en RMN est une évolution prometteuse. Dans ce contexte, la première partie de ma thèse est centrée sur l'analyse de mélanges complexes par RMN avec l'application d'algorithmes de séparation de source à l'aveugle (BSS)[1]. Cette technique, initialement utilisée pour des disciplines telles que les télécommunications et les neurosciences,[2,3] a montré son efficacité pour la décomposition des spectres RMN 1D et 2D.[3] Elle est réalisée à l'aide de corrélations, essentiellement des variations de concentrations, détectées sur une série de données, ce qui permet l'extraction des spectres des composés du mélange.[4,5] Plusieurs types d'échantillons, mélanges synthétiques de terpènes et d'autres, sont utilisés pour évaluer l'efficacité de ces algorithmes.

La deuxième partie de ma thèse concerne deux développements qui peuvent être considérés comme des projets annexes : (i) la mise en œuvre et l'évaluation du filtre $T_{1\rho}$ en métabolomique pour supprimer les signaux des macromolécules qui ne sont habituellement pas pris en compte ; (ii) l'application de la méthode dite « Maximum Quantum » (MaxQ RMN) pour simplifier les spectres RMN de mélanges résultant de la réaction de complexes enzymatiques.

<u>Mots clés</u> : analyse de mélanges, traitement du signal, spectroscopie de résonance magnétique nucléaire

Application and development of NMR methods and signal processing for the analysis of small molecule mixtures

Main project:

Decomposition with signal processing: Blind Source Separation

X + Y → X, Y

Synthetic samples

Source 1
Source 2
Source 3
Source 4
Mixing

SUCROSE

LIMONENE   NEROL   α-TERPINOLENE   β-CARYOPHYLLENE
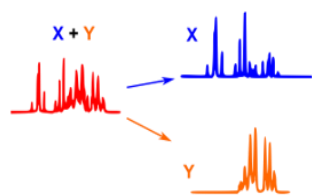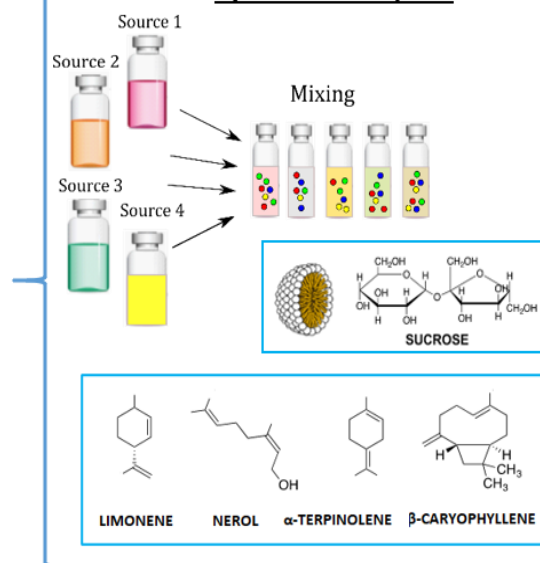
Side projects:

Relaxation filters to remove broad macromolecule signals

Maximum Quantum NMR for enzyme systems

Real samples

Serum / Heart / Liver

Enzymatic systems
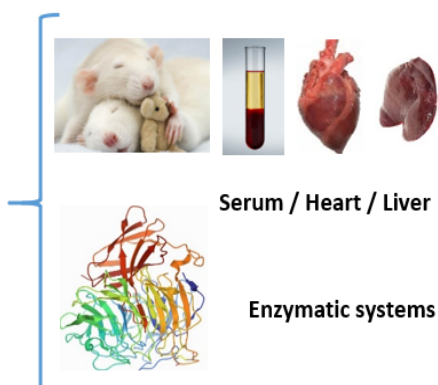
Figure 1 *Graphical summary of this Ph.D. thesis.*

I. Literature Review

I.1. Introduction to NMR spectroscopy

Nuclear magnetic resonance (NMR) is a spectroscopy that exploits the magnetic properties of nuclear spins of a molecule when placed in a magnetic field and allows the identification of different nuclei, such as ^{1}H, ^{13}C, ^{31}P, ^{15}N, according to their characteristic nuclear resonance frequency. Indeed, for each spin moment $\vec{I}$ is associated a magnetic moment $\vec{\mu}$ collinear with $\vec{I}$ which is: $\vec{\mu} = \gamma\hbar\vec{I}$. Here, γ is the gyromagnetic ratio of the considered nucleus, expressed in rad T^{-1}·s^{-1} and $\hbar$ is Planck's constant divided by 2π (where $\hbar$= 1.054 x10^{-34} J·s). All atomic nuclei characterized by a non-zero spin moment $\vec{I}$ are observable by NMR. The proton has a spin I = ½ and it is the first and the most recorded nucleus in NMR owing to its high gyro-magnetic ratio, and natural abundance of more than 99,98 %. Moreover, ^{1}H NMR is widely exploited because most of the known molecules contain hydrogen. For this nucleus, the spins can adopt two states m=+½ and -½ corresponding to two possible orientations of the magnetic moment $\vec{\mu}$. In the presence of a static magnetic field $\overrightarrow{B_0}$, the spins polarize along the axis defined by the direction of $\overrightarrow{B_0}$, either parallel or antiparallel, with two distinct energy levels. It is then possible to induce a transition between these two energy levels if the sample is subjected to an electromagnetic wave of frequency ν_0 obeying the Bohr relation $\Delta E = h\nu_0$, where ν_0 is the Larmor frequency (or resonant frequency) which is written:

$$\nu_0 = \frac{\gamma B_0}{2\pi}(1 - \sigma) \qquad (1.1)$$

σ, is a dimensionless number, called shielding constant, which reflects the shielding effect of the electronic procession on the field $\overrightarrow{B_0}$, as it is perceived at the level of the nucleus. It is of great interest as an analytical tool for NMR since it is specific to a chemical environment and is the origin of the so-called chemical shift.
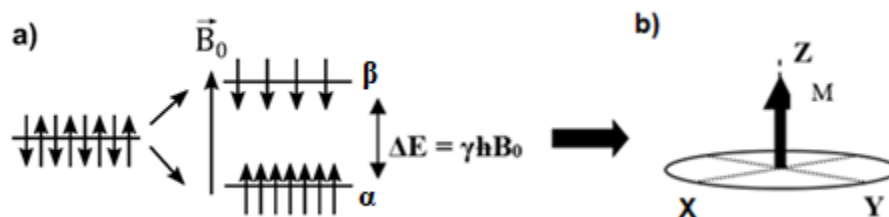
Figure 2 a) *Difference of energy levels of a ½ spin by the application of a static magnetic field $\vec{B}_0$. The parallel orientation is the most stable state noted α.* *b)* *a slight excess of the population corresponding to the parallel orientation to the $\vec{B}_0$ field results in an overall magnetization M over the entire sample, aligned at the equilibrium, along the z-axis.*

As shown in **Figure 2.a**, the most stable state, with the lowest energy, denoted α, corresponding to an orientation parallel to the magnetic field, is the most populated. The overall balance in a sample gives rise to a total nuclear magnetization $\overrightarrow{M_0}$ aligned along the longitudinal axis denoted z, collinear with the magnetic field $\overrightarrow{B_0}$ (**Figure 2.b**). It is a macroscopic quantity, which can therefore be modeled using classical mechanics. However, the nuclear magnetization $\vec{M}$ is not observable when it is parallel to $\overrightarrow{B_0}$. Another radiofrequency field $\vec{B}_1$ perpendicular to $\overrightarrow{B_0}$ must be applied to move it away from this equilibrium position. The equations of classical mechanics show that the magnetization vector $\overrightarrow{M_0}$ is then animated by a precession motion at a frequency which is known as the Larmor frequency (**Equation 1.1**). An NMR signal is obtained upon return to equilibrium: the Free Induction Decay (FID) in the time-domain. The obtained signal is complex and contains several frequencies. By applying a Fourier Transform (FT), we obtain an interpretable spectrum in the frequency-domain, as reported in **Figure 2**, from which information such as chemical shifts (in ppm), scalar coupling, and integrated amplitudes of peaks can be measured.
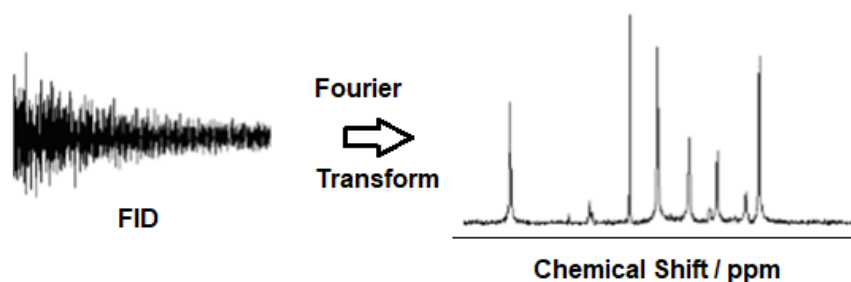
Figure 3 *NMR spectrum in frequency-domain obtained after the Fourier Transform of the detected signal FID in time-domain.*

NMR spectroscopy is a relatively rapid, high throughput, robust and reliable technique that presents a high reproducibility and repeatability, and requires minimal and non-destructive sample preparation. However, it is an approach with a weak sensitivity, for example, compared to Mass Spectrometry (MS). However, the sensitivity can be improved by signal accumulation (repeating the number of scans, using longer analysis times), using higher magnetic fields or/and cryogenic probes. During this thesis, the NMR analyses were performed at high field spectrometers, mainly with a 600 MHz (14.1 Tesla) instrument.

I.1.1 Multidimensional NMR Techniques

Before introducing multidimensional NMR experiments, some technical points have to be detailed describing how one-dimensional (1D) NMR spectrum is obtained for a targeted nucleus, such as proton ^{1}H. The simplest measurement scheme, such as in a 1D ^{1}H experiment (**Figure 4**), is composed of the so-called hard pulse which is a very short (about 10 μs), rectangular, non-selective pulse, performing a 90° ($\pi/2$) tilt angle for the magnetization $\overrightarrow{M}$.
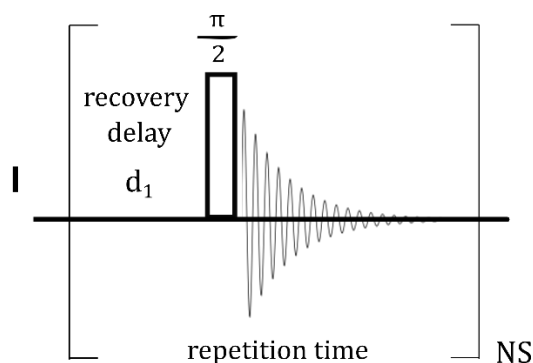
Figure 4 *The simplest one-dimensional pulse sequence for a nucleus I with a relaxation delay to let the magnetization recover, a hard pulse and the FID detection. This measurement scheme can be repeated for a number of scans (NS).*

The measurable signal (FID) is then detected and acquired for a few seconds (*e.g.* 2 s). The signal is accumulated over a repeated number of scans (NS) during which the experiment is repeated in order to obtain a sufficient SNR (Signal-to-Noise Ratio), lengthening the experimental duration. SNR value increases with the square root of NS (*e.g.* if *snr* is the SNR for 1 scan, then the SNR for NS is *snr* x $\sqrt{NS}$). The NS should be a compromise between sensitivity and experimental time. The sum of the repetition time of each scan, in fact, leads to the total duration of the experiment. Performing one-dimensional quantitative experiments requires that the magnetization is completely recovered on the longitudinal axis before each scan.

Since the excitation part is not always performed by a single hard pulse, NMR multi-pulse sequences have been developed to provide different information, using radio-frequency (RF) pulses, delays, and magnetic field gradients.

In this context, multi-dimensional pulse sequences can be introduced as part of multi-pulse experiments. In a two-dimensional (2D) experiment the signal is recorded as a function of two variable times, t_1 and t_2, and then the double Fourier transform produces a 2D spectrum as a function of two frequency-domain variables, F_1 and F_2. In practice, a 2D spectrum is recorded as a series of 1D experiments collected with different incremented t_1 to build the second dimension, as shown in **Figure 5**. Repeating the pulse sequence, that is increasing the t_1,

provides an experimental time by the product of the repetition time, the NS, and the number of increments.
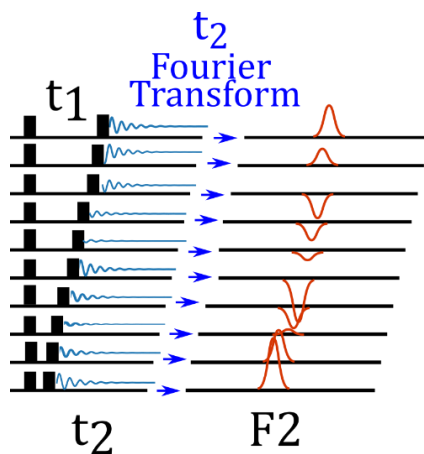
Figure 5 *The collection of the series of 1D experiments with t_2 along t_1 in 2D experiments.*

The general scheme of any 2D experiment consists of four blocks as shown in **Figure 6**. The first, the preparation which is defined by a delay allowing the relaxation of the system and the excitation of a population of spins, but may also have solvent suppression pulses. The second, the evolution time t_1, is the incremented delay for mapping of chemical shifts, during which the system evolves. This evolution can take place under the effect of the chemical shift and/or the couplings between nuclei observed in the F_1 dimension. The third is the mixing time during which through-bond (J_{HH}) long range couplings take place. This transfer takes place by different mechanisms, scalar coupling, dipolar interaction, *etc*. Finally, the experiment is finished with the detection time t_2 for the acquisition of normal FID (**Figure 6**).
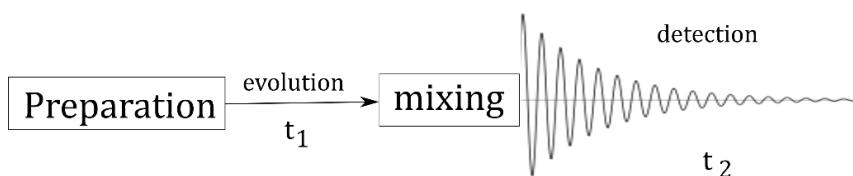
Figure 6 *Basics of any 2D NMR experiment: general scheme.*

Resolution in the indirect dimension t_1 is very important. Higher resolution in direct t_2 dimension costs little time, but the higher resolution in t_1 adds time directly to the total

experimental time (for example, doubling the number of t_1 leads to an experiment that is twice as long).

2D NMR experiments can be of two types: homonuclear and heteronuclear. Each type can provide either through-bond (COSY-type) or through-space (NOESY-type) coupling information.

I.1.1.1. Homonuclear NMR Experiments

The first and simplest two-dimensional NMR experiment is the homonuclear proton-proton COrelation SpectroscopY (COSY).[7] COSY spectra show connections between neighbour protons through non-diagonal peaks or cross peaks. It generates a 2D map which mostly shows cross peaks from geminal (^2J couplings, where two links separate two protons) and vicinal (^3J couplings) groups only. The simplest pulse sequence is constituted by two 90° pulses separated by the evolution period t_1 as shown in **Figure 7**.
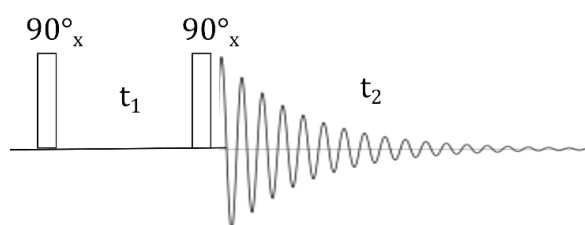
Figure 7 *Basic 2D Homonuclear proton-proton COSY experiment with the pulse sequence.*

Two-dimensional ^1H-^1H TOCSY (TOtal Correlation SpectroscopY) is a homo-nuclear experiment, generally used for protons.[7] It gives a spectrum similar to that one of COSY, with additional correlations between all the protons within a spin system, in which coupling between two spins is indicated by the presence of a cross-peak multiplet. In TOCSY, the net magnetization transfers proceed from one spin to another through space even without direct coupling. If a spin A is coupled to spin B, and B is coupled to spin C, then in this kind of

spectrum we will see a cross peak between A and C, and also between A and D, as reported in **Figure 8**, even if there is no direct coupling between these two spins.
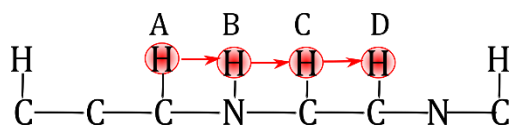
Figure 8 Magnetization transfer from one spin to another even without direct coupling. If a spin A is coupled to spin B, and B is coupled to spin C, also a cross peak appears between A and C, and also between A and D.

The key part of this experiment is the period of isotropic mixing, during the mixing time τ_{mix}, which forms the mixing period as shown in **Figure 9**. The difference from COSY is that the second 90° pulse is replaced by the spin-lock period which uses continuous irradiation along a transverse axis.
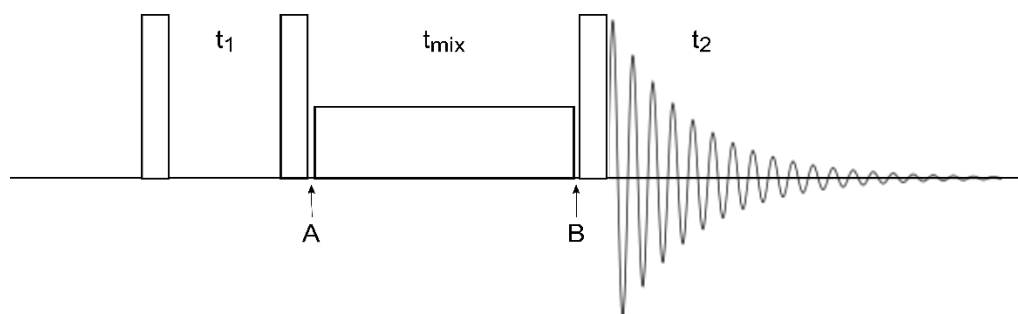
Figure 9 Pulse sequence for the TOCSY experiment. During the mixing time there is the isotropic mixing which transfers magnetization between spins which are connected via a network of couplings. It is achieved by a specially designed multi-pulse sequence such as DIPSI-2. It is arranged that only z-magnetization present at points A and B contributes to the spectrum.

TOCSY experiment is extremely useful to identify networks of coupled spin. It is used for peptides analysis,[8] since almost all of the side chains form one contiguous spin system, allowing the assignment of amino acids.[8] However, for most organic compounds, the spin systems are isolated into several fragments by the presence of heteronuclei or non-protonated carbons. In such cases, additional NMR experiments or the use of heteronuclear-edited ^{1}H TOCSY experiments (such as 2D HSQC-TOCSY[9] or 3D heteronuclear TOCSY[10]) are required to complete the assignment.[8]

In this work, the assessment of the application of Blind Source Separation algorithms on TOCSY data sets is performed in view of future potential applications.

I.1.1.2. Heteronuclear NMR Experiments

In the previously introduced NMR experiments, only one nucleus at the time was observed. Two-dimensional Heteronuclear Single Quantum Coherence (HSQC) sequence is the benchmark hetero-nuclear experiment allowing to highlight interactions between different nuclei such as ^{1}H-^{13}C correlations providing significant resolution and discrimination of resonances since the correlation peaks are carried by the chemical shifts of two different nuclei. This experiment is of great interest when the compounds to be studied have very similar structures with overlapping peaks.[11] Indeed, the large spectral dispersion of ^{13}C (0 to 250 ppm, against 0-15 ppm for proton) allows an increased spectral resolution. Cross peaks here indicate proton - carbon chemical bounds, combining the good sensitivity of ^{1}H with the good resolution of ^{13}C nucleus.

In the HSQC experiment, the preparation period consists of a delay allowing the return to equilibrium of the magnetization and an Insensitive Nuclei Enhanced by Polarization Transfer (INEPT) block which transfers the polarization from the protons to the carbons.[12] **Figure 10** shows the basic HSQC pulse sequence built using two INEPT refocusing blocks. The INEPT composes the first part of the HSQC pulse sequence, transferring the magnetization from ^{1}H to ^{13}C, providing the evolution of magnetization of carbon spins during the period t_1 under the effect of chemical shift.
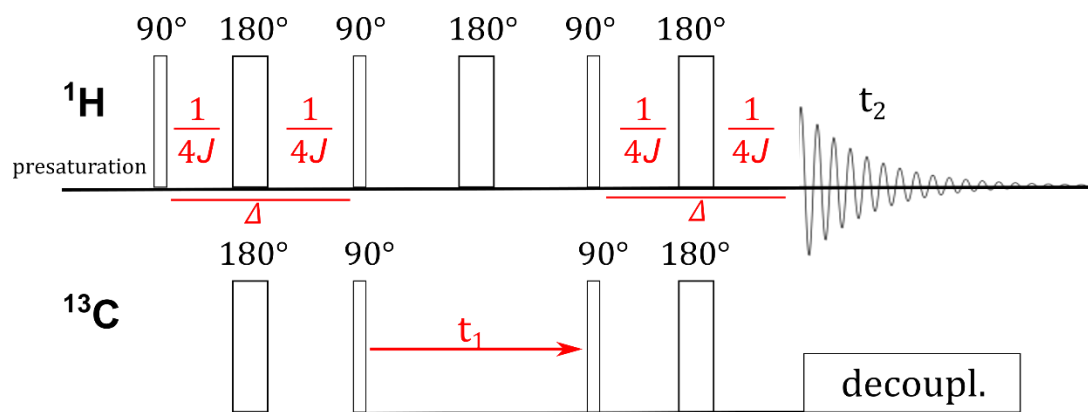
Figure 10 The simplest 2D HSQC pulse sequence, with water presaturation, for ^{1}H-^{13}C correlations with decoupling during acquisition. The INEPT delay is $\Delta = 2\frac{1}{4J_{IS}} = \frac{1}{2J_{IS}}$.

INEPT transfers the nuclear spin polarization from sensitive spins, *e.g.* ^{1}H, to less sensitive ones, such as ^{13}C, that has a low gyromagnetic ratio and low natural abundance, in order to increase the sensitivity of the signal. During the evolution period, only the chemical shifts of ^{13}C are expressed during the incrementable time t_1 thanks to a 180° ^{1}H pulse in the middle of this period which refocuses the ^{1}H-^{13}C couplings as shown in **Figure 10**. This gives a proton-decoupled ^{13}C spectrum in the indirect F_1 dimension. The mixing period allows the magnetization of carbons-13 to be transferred to protons with the same INEPT block. In the end, the ^{1}H NMR signal is detected in the presence of a ^{13}C decoupling, and therefore, during t_2, the proton-carbon couplings are eliminated from the spectrum.

In this thesis we used the standard "HSQCETGPSI" (according to the Bruker nomenclature) pulse sequence which is a phase-sensitive gradient HSQC using Echo-AntiEcho quadrature with sensitivity improvement.[7,13]

The direct relation between signal surface or volume of the peaks and concentration is partially modified in conventional 2D NMR spectra, where the factors which affect the quantitative response are much more complex than with a simple one-dimensional excitation-acquisition scheme of **Figure 4**. The multidimensional equivalent of the integral, "the hyper-volume element", becomes proportional to the transfer functions which take into account all interactions

contributing to produce the n-dimensional cross-peaks (J couplings, chemical shifts, relaxation, *etc.*), and depending on the timing and architecture of the pulse sequence. In this setting, the coefficient of proportionality between the signal and the concentration is different for each peak.[14]

In this framework, the problem in HSQC is the non-quantitative response due to the dependence of the correlation peak volume (V_C) to be integrated on the efficiency of the magnetization transfer of the two INEPT steps, which is proportional, among other things, to the heteronuclear coupling constant (J_{IS}). The peak volume varies according to:[15]

$$V_C \propto \sin^2(\pi \Delta J_{IS}) \qquad\qquad (1.2)$$

where $\Delta = \frac{1}{2J_{IS}}$ is the polarization transfer delay of the INEPT block, corresponding to two times $\frac{1}{4J}$. In the classical HSQC, Δ is adjusted to the average value of one-bond $^{13}C-^{1}H$ coupling constant $^{1}J_{CH}$ of 145 Hz corresponding to 3.45 ms to maximize the transfer efficiency of magnetization.[15] Different coupling values can be maximized by changing Δ, but if there is a range of coupling constants, there is not one single value that would maximize the intensity of all of the resonances. The transfer efficiency will differ between peaks in the HSQC spectrum, with the consequence that the peak integrals will not be quantitative. Several strategies were proposed to overcome this issue. For example, in Q-HSQC, proposed by Heikkinen *et al.*, the $^{1}J_{CH}$-dependence of correlation peak volumes, so the transfer efficiency is compensated by summing signals from experiments with different values for the INEPT delays.[15] The relatively uniform response of V_c is achieved by adding together data from experiments with four values of Δ of 2.94 and 5.92 ms in a 3:1 ratio to average the dependence on $^{1}J_{CH}$ at less than 2% for $^{1}J_{CH}$ between 100 and 240 Hz.[16] The resulting spectrum, however, will have less sensitivity with a volume 25% lower than the conventional HSQC spectrum when actual $^{1}J_{CH}$ is close to $^{1}J_{CHexp}$.[15] In the Q-HSQC, if a proton has many J_{HH} couplings, the relatively long coupling

evolution period during polarization transfer INEPT periods, could lead to distort the line shape of the correlation peak, compromising the accuracy of the result and causing problems in the cross-peak integration.[17,18] There is the combination of signals acquired using four different INEPT delays, consequently, the experiment requires four times as many scans as the HSQC experiment to achieve the same resolution in the indirect dimension. This problem was addressed later in the Quantitative Carr-Purcell-Meiboom-Gill (CPMG)-adjusted HSQC (Q-CAHSQC) experiment, using CPMG pulse trains by replacing the constant-time INEPT periods with the constant-time CPMG-INEPT steps, allowing to suppress the evolution of J_{HH}, and correct the cross-peaks shapes.[18] Quick Quantitative HSQC (QQ-HSQC) has the quantitative properties of the Q-HSQC experiment but only requires as many scans as a conventional HSQC experiment. The signals corresponding to different INEPT delays are acquired simultaneously from different parts of the sample. Koskela et *al.*[18] introduced later optimized the RF pulses using quantitative, offset-compensated, CPMG-adjusted HSQC (Q-OCCAHSQC) experiment, with improvement to the offset performance, covering the carbon chemical shift range.

In this thesis, applying the BSS on HSQC data sets acquired with the same pulse sequence (the standard HSQCETGPSI), we look at the samples in a relative way. The changes in the volume intensities between the mixtures are only due to the variations in concentrations from one sample to another since we used the same INEPT delay for the HSQC ($\Delta = \frac{1}{2J_{IS}} = 3.45\ ms$), and the same parameters: recovery delay d_1, number of scans and dummy scans, acquisition time...*etc.*

I.1.1.3. Accelerating the multidimensional NMR

An NMR experiment may last from a few seconds to several hours, even some days for some multidimensional experiments. Furthermore, long acquisition times are not suitable for large batches of samples or reactive molecules.

To save experimental time, there are two very common approaches that were not used in this work but that are described here. Multidimensional NMR experiments can be acquired with a single scan using UltraFast (UF) multidimensional acquisition techniques, limiting the duration of multidimensional NMR experiments into the lifetime of reactive systems, for example.[19] UF technique involves a heterogeneous initial excitation with a single axis B_0 field gradient, introducing artificial spectral heterogeneities within the sample, resulting in inhomogeneous spectral distribution. This leads to view the sample as composed of independent sub-ensembles, where their evolution can be manipulated selectively.[19]

The other technique is Band-Selective Optimized-Flip-Angle Short-Transient (SOFAST) which combines the advantages of longitudinal-T_1 relaxation enhancement or optimization and Ernst-angle ($0° <\alpha< 130°$) excitation.[20] [21] This is useful to accelerate 2D experiments, by using very short inter-scan delays, according to the Ernst Angle strategy.

I.1.1.3.1. Non-Uniform Sampling (NUS)

The acquisition time of a 2D spectrum largely depends on the number of points, the increments, acquired in the indirect dimension (TD_1, or time-domain points in t_1, which represents the size of FID). It depends also on the duration of the acquisition of the direct spectrum, called recycling time or repetition time, *TR,* and the number of scans, *NS* (introduced in paragraph I.1.1.1). Based on the Nyquist sampling theorem, there is a limited flexibility on the choice of the size of the data matrix to ensure a suitable resolution and a wide enough frequency range, linked respectively to the number of the acquired indirect points and the sampling increment, Δt_1. Resolution in the indirect dimension F_1 must be large enough to resolve the correlations of interest, but the higher resolution in F_1 implies a longer time for the construction of the indirect dimension and thus a longer acquisition time. Resolution and spectral width are obviously more critical in the context of mixture analysis.

The acquisition parameters of an NMR experiment – sampling rates and acquisition times in every dimension – are used to define the so-called data matrix, which tabulates the NMR signal intensity as a function of frequency in every dimension. With this in mind, the acquisition of a smaller data matrix is a solution to reduce the nD experiment duration, with the desired resolution recovered by signal processing. A practical procedure consists in skipping acquisition points acquired in the indirect dimension t_1 from the list of the full experiment forming the full Nyquist grid. This leads to the so-called Non-Uniform Sampling (NUS) method (**Figure 11** and **Figure 12**).
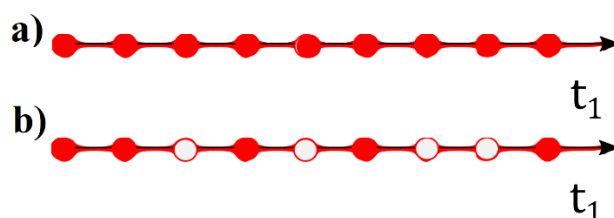
Figure 11 a) Uniform Sampling in a 2D experiment, in which all the indirect data points are collected, from t_1=0 to t_{1max}. b) 2D NUS experiment with only a subset of complex points are collected (in red) (non-collected points are in white).
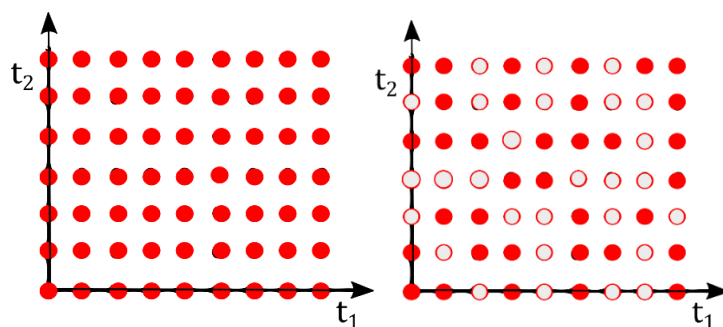
Figure 12 Uniform sampling of points in 3D experiment (on the left-hand side). Distribution of complex points in a random NUS grid for a 3D experiment. Only a fraction of t_1/t_2 pairs are collected (on the right-hand side).

Nyquist sampling theorem dictates that the sampling interval must be at least 1/DW. In NMR, dwell time (DW) is defined as the time between two successive points during data acquisition in the FID, also called the sampling rate which is also equal to $\frac{1}{2SW}$, with SW the Spectral Width in Hz. Time-domain measurement of the FID consists in sampling the signal at a series of discrete intervals, where these intervals are kept uniform (thus uniformly sampling, US).[22] According to the classical Nyquist-Shannon sampling theorem, sampling at a constant rate,

which is equal or larger than the spectral width, is the necessary condition of the reconstruction of a spectrum. NUS acquires only a subset of these data points in a random manner that is sufficient for obtaining the exact reconstruction of the spectrum. NUS percentage of collected t_1 points can be different. It is defined as the amount of sparse sampling equal to the ratio between the number of keep points and the total points. For example, with 25% of NUS points, the compression factor is equal to 4 (**Figure 13** with t_1 keep points for NUS 25% at the top, 12,5% in the center, and 6,25% at the bottom).
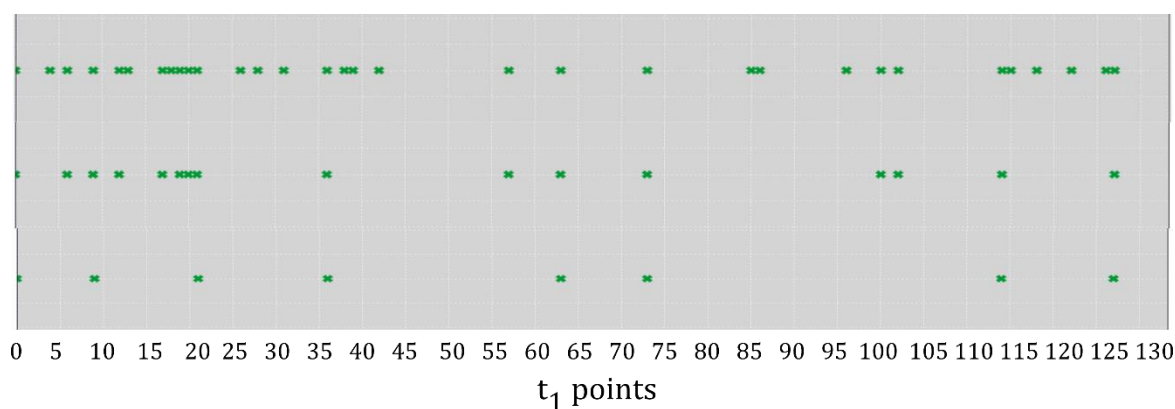
Figure 13 The distribution of t1 complex points recorded in the indirect dimension in a 2D NUS experiment from Topspin, with the 25% on the 128 uniformly sampled points (first line), 12,5% (second line) and 6,25% (third line).

The search for algorithms leading to a faithful compression/reconstruction has been the subject of numerous researches and theoretical developments, from which the most commonly employed NMR data processing approaches were adapted.[23] The quality of processing for NUS data is assessed based on the existence of several artefacts, including baseline and peak shape distortion, aliasing (including "reconstruction noise"), and signal loss. The presence of such artifacts is clearly an even bigger issue for mixture analysis when the artifacts originating from the most intense peaks have intensities comparable to or larger than the resonances of the minor components. Several processing methods are available such as Maximum Entropy (MaxEnt),[24, 25,26,27] Multi-Dimensional Decomposition (MDD-NMR),[28,29,30] Compressed Sensing or compressive sampling (CS),[31] Multidimensional Fourier Transformation (MFT).[32,33,34]

Incomplete sampling multidimensional techniques rise essentially two questions: firstly, the limit of compressing the data acquisition set without losing accuracy and, secondly, the selective signal modifications according to the peak multiplicity. An important change in line-shapes was reported by increasing the number of points dropped by NUS while preserving the same resolution. For mixtures, this is still a matter of investigation, and it was recently shown that the gain in time, by increasing the compression ratio with the same resolution, corresponds to a loss in signal.[35,36] Le Guennec *et al.* demonstrated that for HSQC experiments, the integrated values of the peaks with low intensity dropped below 50% when NUS was employed.[35]

I.1.2. NMR Dynamic Parameters

In this section, we will approach the dynamic aspect of NMR, which consists in relating the relaxation and diffusion parameters to molecular dynamics. We will present the relaxation parameters that we actually used in this thesis and we will define the self-diffusion coefficient which allows access to translational movements. We will devote a part to the experimental determination of the relaxation parameters and the self-diffusion coefficients. The spin relaxation and the diffusion coefficient determined by NMR depend to a large extent on molecular mobility and can provide fine structural information such as interatomic distances or more global such as the size of molecules or the viscosity of the medium.[37–39]

I.1.2.1. Relaxation Parameters

NMR relaxation involves the processes allowing a nuclear spin system to return to its equilibrium state after being removed from it, achievable by various disturbances. It can be understood by a phenomenological approach using Bloch's equations describing the evolution

of nuclear magnetization as a function of time.[37] It should be pointed out that this approach which considers a global macroscopic quantity is useful for understanding relaxation but it is unsuitable for describing spin systems containing more than two states. In this case, quantum mechanics gives a complete description by no longer considering macroscopic quantities but rather the different spin states.

As described previously, the equilibrium state corresponds to a magnetization $\overrightarrow{M_0}$ aligned along the z-axis which is collinear with the magnetic field $\overrightarrow{B_0}$ ("longitudinal direction"). When the nuclear magnetization M_0 is brought out of its equilibrium position by a radiofrequency (rf) pulse, two different processes bring the system back to its equilibrium state: the longitudinal relaxation which describes the reconstruction of the magnetization along the longitudinal z-axis and the transverse relaxation which rules the vanishing of components of the magnetization in the transverse plane (x, y).

I.1.2.2 Measurement of the longitudinal relaxation time T_1 by Inversion-Recovery

In general, any measurement of relaxation time begins with a disturbance of the system to bring it out of its equilibrium state, followed by an evolution period during which the magnetization returns to equilibrium according to the relaxation process.[40] The state of the system is usually observed by an rf read pulse ($\pi/2$ for example). The experiment is repeated for different values of the evolution time in order to follow the temporal evolution of the magnetization to extract the relaxation time(s). There are different methods of measuring T_1, they are differentiated by the initial disturbance of the system which can be one or more rf pulses. In this thesis, we used the inversion-recovery experiment (**Figure 14.a**) which begins with an inversion π pulse that changes the nuclear magnetization from $+M_z$ (equilibrium) to $-M_z$. It is followed by an evolution delay during which the magnetization evolves and returns

to its equilibrium position along the *z*-axis, passing through the zero value (**Figure 14.b**), according to the following **Equation 1.3**:

$$M_z(\tau) = M_z(0) \left(1 - 2\, e^{-\frac{\tau}{T_1}}\right) \qquad (1.3)$$

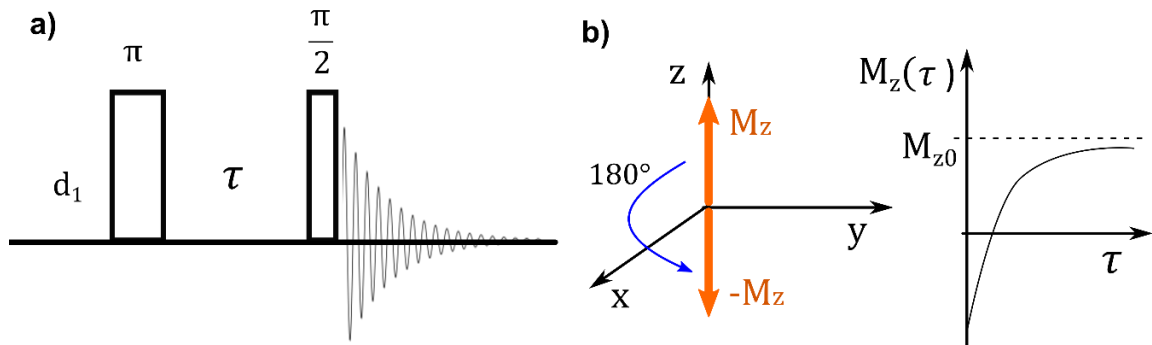The state of the magnetization is read by a $\pi/2$ pulse (**Figure 14.a**).

Figure 14 a) Inversion-recovery sequence for measuring T1; b) Evolution curve of the amplitude of the NMR signal as a function of the evolution time.

I.1.2.3 Measurement of relaxation times in the transverse plane, T_2 and $T_{1\rho}$

I.1.2.3.1 T_2 Measurement by the Carr-Purcell-Meibomm-Gill (CPMG) sequence

The Carr-Purcell-Meibomm-Gill (CPMG) technique is based on the Hahn sequence also known as "spin echo" which allows, among other things, to overcome the contribution of the inhomogeneity of the magnetic field.[41,42] Indeed, the inhomogeneity of the magnetic field B_0 leads to a dispersion of the precession frequencies (proportional to the effective field in each position in the sample). As a result, the decay of the free precession signal is faster with a constant T_2^* lower than the true transverse relaxation time T_2. As with any relaxation measurement, maximum disturbance should be applied. In the case of transverse relaxation, it consists of a $\frac{\pi}{2}$ pulse which brings the magnetization into the *x, y* plane. The application of a π pulse at the midpoint of the echo delay τ cancels out the effect of the inhomogeneity of the magnetic field and "refocus" all precessions of the nuclear magnetization. The decay of the

magnetization after the delay τ is only due to the transverse relaxation T_2. The amplitude of the

signal measured at the end of the delay τ is given by **Equation 1.4**:

$$M_{xy}(\tau) = M_{xy}(0) \; e^{-\frac{\tau}{T_2}} \qquad\qquad (1.4)$$

However, the impact of molecular diffusion on the spin-echo experiment has been demonstrated

and leads to a faster signal decrease if the static magnetic field is not homogeneous.[41,42] Indeed,

after the π pulse and the second half of the echo delay, the magnetization returns to the starting

position (along the *y*-axis for example) provided that its precession is done with the same

frequency, this implies that the molecules have remained nearly in the same position during τ.

Nevertheless, during the refocusing process the molecules undergo a translational movement,

and, if the magnetic field, therefore the precession frequency, differs significantly from one

position to another in the sample, an additional attenuation appears, proportional to the self-

diffusion coefficient *D*.

Despite the sufficient homogeneity of the magnetic fields of recent instruments, the CPMG

sequence is preferred for the measurement of T_2 and for the application of relaxation filters. It

consists of replacing the single π pulse of the Hahn method with a π pulse train (**Figure 15**)

which enables to overcome the effects of molecular diffusion, even in the presence of internal
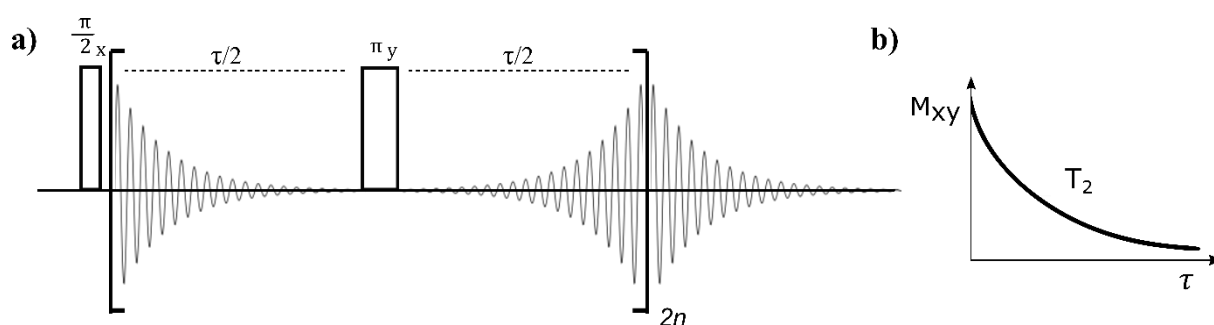
gradients.

Figure 15 *Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence to measure T_2 relaxation time **(a)**, less affected by translational diffusion, using the intensity=f(τ) to derive T_2 **(b)**.*

Regarding the implementation of CPMG, the length of the pulse train is increased (interval τ remains constant), by increasing the number of loops $2n$. The transverse relaxation time T_2 is then determined by adjusting the equation:

$$M_{xy}(2n\,\tau) = M_{xy}(0)\, e^{-\frac{2n\tau}{T_2}}$$

A more recent approach is the Periodic Refocusing of J Evolution by Coherence Transfer, PROJECT. It is based on the perfect echo approach by Takegoshi *et al.*,[43] consisting in a CPMG with a 90°$_y$ refocusing pulse at the midpoint of a double spin-echo, to suppress the homonuclear J modulation. The presaturated sequence ([presat $- 90° - (\tau - 180° - \tau - 90° - \tau - 180° - \tau$ $-)_n$]) from Aguilar et *al.*,[44] is reported in **Figure 16**.
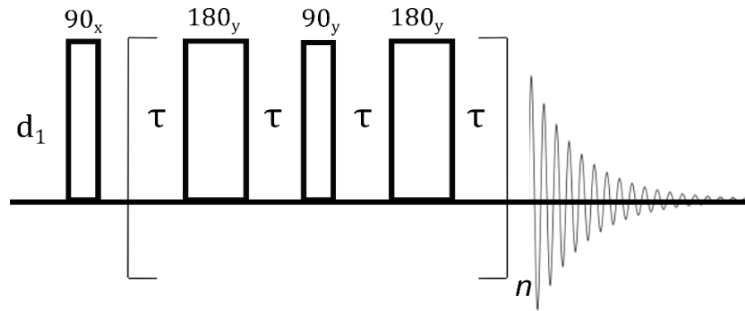
Figure 16 Presaturated PROJECT sequence ([presat – 90° – (τ – 180° – τ – 90° – τ – 180° – τ –)$_n$]) from Aguilar et al.[44]

I.1.2.3.2 Measurement of the relaxation time in the rotating frame $T_{1\rho}$

When the studied systems include scalar couplings, *J*, which was the case of the systems studied during this thesis, the correct measurement of T_2 is hampered due to the modulation of the amplitude of the echoes due to the *J* couplings. The increasing of the π pulse frequency by reducing the inter-pulse delays $\left(\frac{\tau}{2}\right)$ leads to measurements that are increasingly insensitive to diffusion phenomena and to the modulation by *J* couplings. An extreme reduction of these delays would lead to approaching a locking of the magnetization along a transverse axis, *y* for

example. This situation is similar to a $T_{1\rho}$ measurement method which allows to determine the relaxation time in the rotating frame.

The sequence for measuring $T_{1\rho}$ is described in **Figure 17**: a $(\pi/2)_x$ pulse brings the nuclear magnetization along the *y*-axis of the rotating frame. It will be locked by the continuous application of a B_1 *rf* field (stationary along the *y* axis of the rotating frame, called "*spin-lock*") for a time τ with an appropriate amplitude which avoids the shift between the carrier frequency and the resonance frequencies ("offset effect"). The evolution of the magnetization takes place without precession and $T_{1\rho}$ is obtained by adjusting **Equation 1.5** which rules the disappearance of the signal as a function of time τ of the spin-lock (SL).

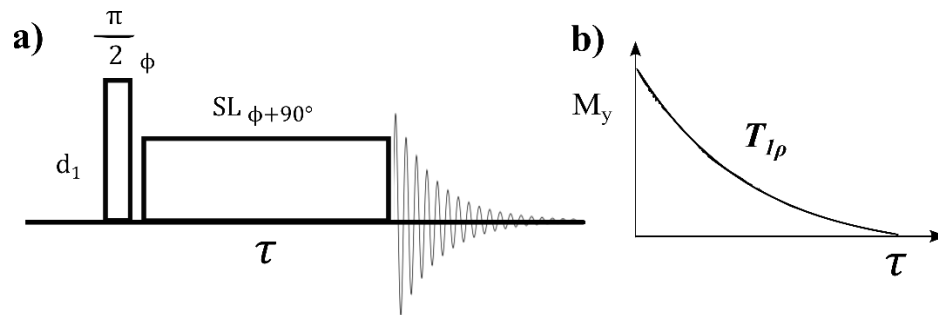$$M_y(\tau) = M_y(0)\, e^{-\left(\frac{\tau}{T_{1\rho}}\right)} \qquad (1.5)$$

Figure 17 a) *The basic spin-locking pulse sequence for $T_{1\rho}$ measurement. The initial $\pi/2$ pulse converts the initial longitudinal magnetization into transverse magnetization along the rotating frame x or y axis. The parameter τ concerns the spin-lock length, adjusted according to the sample, but usually set to 100 ms.* ***b)*** *The intensity=f(τ) fitting curve to derive $T_{1\rho}$.*

For the measurement of $T_{1\rho}$, it is suitable to know the spin-lock amplitude expressed generally according to the pulsation $\omega_1 = \gamma B_1$. Indeed, the $T_{1\rho}$ is very sensitive to the amplitude of the *rf* field. It is most often expressed by the frequency $\nu_1 = \frac{\gamma B_1}{2\pi}$ which is the nutation frequency.

I.1.2.4. Translational Diffusion Measurement by NMR

The relaxation parameters that we have described are essentially related and modulated by molecular rotational movements whose characteristic times are of the order of a

microsecond to a nanosecond. Regarding the translational movements, directly quantified by NMR, they can be tracked for relatively long times, from a millisecond to seconds. More recently, the use of long-lived singlet spin order as a means of storing magnetization during the diffusion time has made it possible to follow the translation of molecules for durations of up to tens of minutes in order to measure tortuosity in porous media.[45]

The following in this paragraph implies the assumption that the molecular translation is random and isotropic, which leads to the definition of the self-diffusion coefficient D according to Stokes-Einstein **Equation 1.6**:

$$D = \frac{k_B T}{6\pi\eta R_h} \tag{1.6}$$

with D = diffusion coefficient (m^2/s), k_B = Boltzmann constant (1,380649 x10^{-23} J K^{-1}), T = temperature of the system (K), η = medium viscosity, R_h = hydrodynamic radius of a sphere with equivalent hydrodynamic properties to the diffusing species. What is measured in NMR is actually a root mean square displacement. This diffusion coefficient generally reflects the size of molecules or the interactions they undergo. These features gave rise to several applications, the most popular consists in analyzing solutions containing complex mixtures (artificial analytical chromatography).[46] Later, Morris and Johnson proposed the DOSY (Diffusion Order SpectroscopY) technique, described below, which is a representation of the result of the diffusion experiment according to a two-dimensional map, in the "direct" dimension the spectrum showing the chemical shifts, and in the "indirect" dimension the diffusion coefficients.[47–54]

To measure the diffusion with NMR, magnetic field gradient pulses are used. One of the basic pulse sequences is the Pulsed Field Gradient Spin Echo experiment (PFGSE) (**Figure 18**).
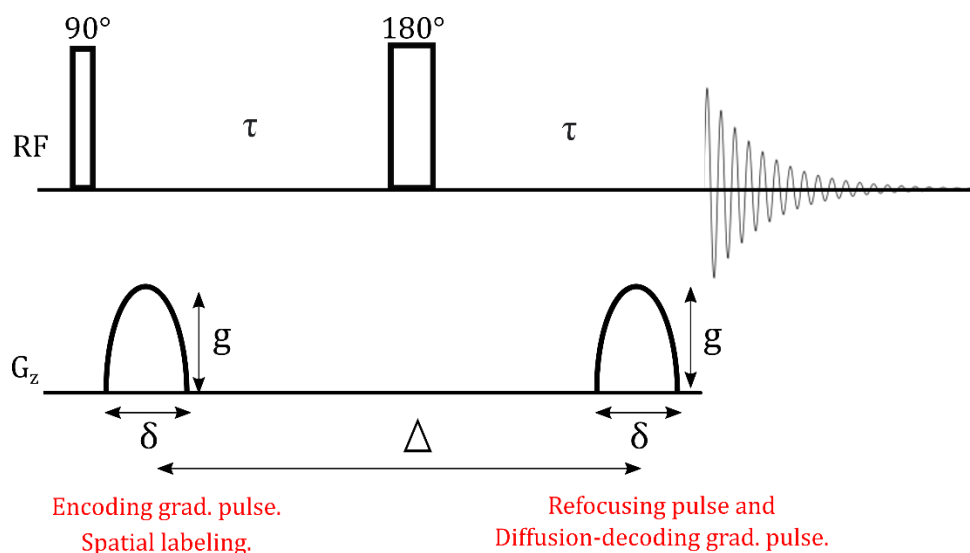
Figure 18 *PFGSE DOSY NMR schema of pulse sequence with δ, the duration of gradient, Δ, the diffusion delay, and g, the gradient strength.*

It is a spin-echo sequence in which two equal gradient pulses of the same intensity and duration δ are integrated into both τ periods. A first magnetic field gradient defocuses the nuclear spins and provides spatial labeling where each molecule is labeled according to its position in the sample. It is followed by a diffusion delay noted *Δ*, during which the molecules diffuse, and therefore change their position. Since the *"encoding"* gradient corresponds to a controlled inhomogeneity of the static magnetic field $\vec{B}_0$ which gives to each spin a position-dependent phase, a second *"decoding"* gradient pulse, equal in duration δ and magnitude (*g*, expressed in G/cm), will refocus the spins (it cancels the effect of the first). This complete refocusing is possible only when the magnetic field experienced by the spins is the same before and after the diffusion delay *Δ,* provided that the molecules have not changed position during this delay. Actually, the molecules move in random translational motion, and the spins that diffuse do not experience exactly the same magnetic field during the second gradient pulse. Only partial refocus occurs, resulting in loss of NMR signal according to the Stejskal – Tanner **Equation 1.7**:[51]

$$I = I_0 exp^{[-D\,\gamma^2\,g^2\,\delta^2\,\Delta']}\ exp^{\frac{-2\tau}{T_2}} \qquad (1.7)$$

where Δ' is the diffusion delay which takes into account the diffusion during the gradient pulses, δ and g are the length and the amplitude of the gradient pulse, respectively, γ the gyromagnetic ratio of the observed nuclei, and D the diffusion coefficient. In practice, a series of experiments is carried out for different values of the magnetic field gradient g, while keeping the diffusion delay Δ constant (**Figure 19**). Thus, the decay due to relaxation $\left(exp^{\frac{-2\tau}{T_2}} \right)$ is the same from one experiment to the next and the decay of the curve is only due to diffusion.

For example, in a mixture of two molecules, the signal intensity of small molecules, with high D, decrease more rapidly (**Figure 19.a**) than that of large molecules with lower D (**Figure 19.b**).
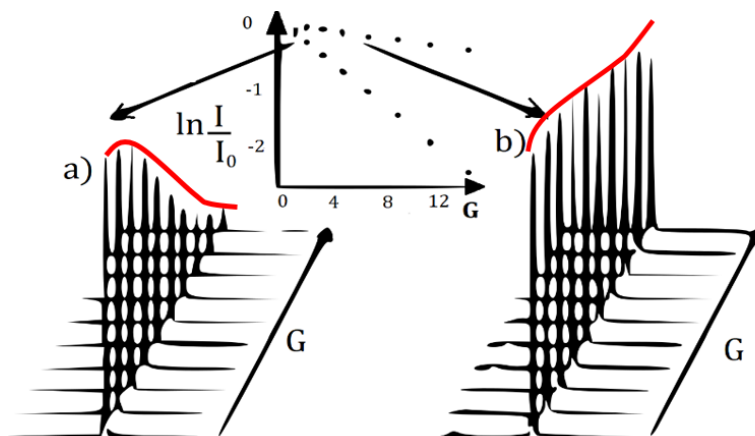

*Figure 19 Signal diffusion decay as a function of gradient strength (G/cm) for a small molecule with a large D **(a)** and for a large molecule which diffuses slower and with a smaller diffusion coefficient **(b)**. Reprinted with permission from Cohen et al., Angewandte Chemie, 44 (2005) 520–554. Copyright 2005 Wiley.[55]*

As mentioned before, Fourier Transform (FT) is the data processing associated with most NMR experiments to transform time-domain signal to the frequency-domain spectra. Exceptions are done when the signal is characterized by multi-exponential decay, for example, in spin relaxation or in the dephasing of the NMR signal associated with the PFGs, where only one dimension is suitable for the FT. From PFG data, DOSY charts are obtained thanks to Laplace Inverse Transform (ILT) (**Figure 20**).[56]
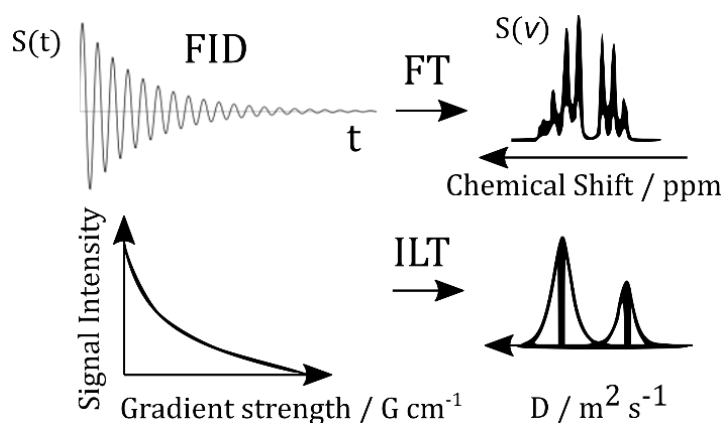
Figure 20 *Comparison of FT and ILT transformation for PFG data. Reprinted by I. Toumi thesis (2013).[57]*

The Fourier transform of the NMR signal and the inverse Laplace transform of the decay of the signals lead to a DOSY spectrum on which the chemical shifts (δ, in ppm) are located on the horizontal axis, while the diffusion coefficients are on the vertical axis (D, in $m^2 \cdot s^{-1}$). All the signals of the same component will be aligned in the horizontal direction according to their chemical shifts and in the vertical direction according to their diffusion coefficients. The limit is that the different compounds of the mixture can be separated as a function of the value of their diffusion coefficients at the condition that spectra do not have very severe overlaps. In **Figure 21**, an example of resolved 2D DOSY chart of an equimolar mixture of toluene, benzyl alcohol, and tetraethylene glycol (TEG) in D_2O (HOD) is presented. Four compounds are present according to the four groups of signals aligned to four different diffusion coefficients (dashed lines in **Figure 21**).[48,58]
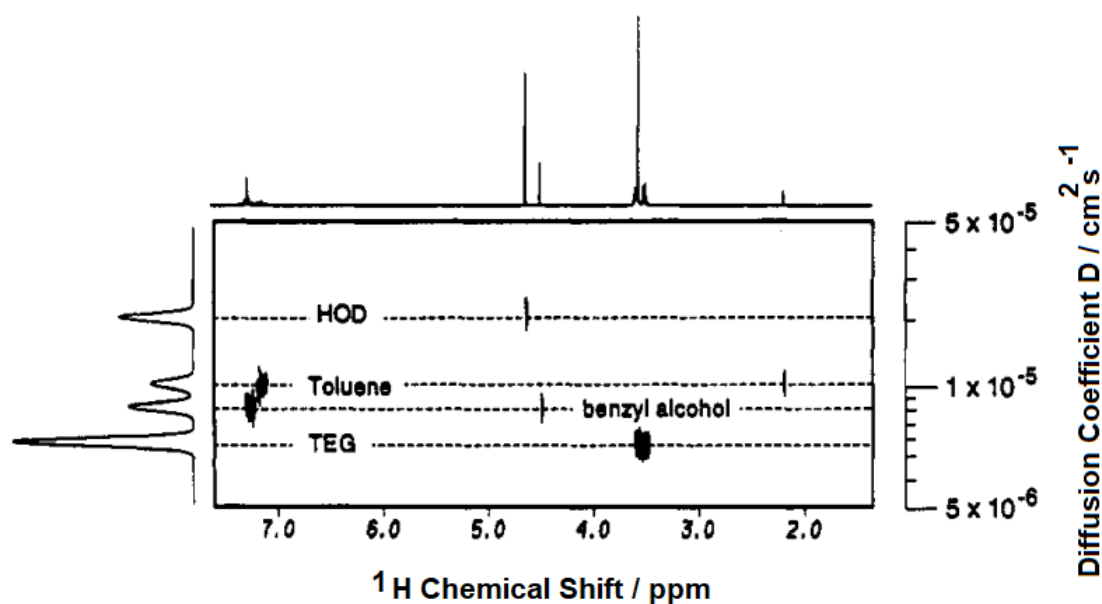
Figure 21 *An example of resolved 2D DOSY spectrum of an equimolar mixture of toluene, benzyl alcohol, and tetraethylene glycol (TEG) in D$_2$0 (HOD). The dashed lines indicate which peaks belong to a same component. Four compounds are present according to the four groups of signals aligned to four different diffusion coefficients. Reprinted with permission from Morris et al., Anal. Chem., 66 (1994) 211–215. Copyright 1994 American Chemical Society.[48]*

For some samples, the attenuation due to transverse relaxation is very large, which can interfere with a good measurement of the self-diffusion coefficient. Therefore, the most used approach is the stimulated echo sequence (PGSTE)[52] which consists of storing the magnetization along the *z*-axis during *Δ*. Consequently, the additional attenuation is due to longitudinal relaxation (*T$_1$*) which is longer than the transverse relaxation (*T$_2$*).

I.1.3. Mixture Analysis by NMR

Analytical chemistry techniques for mixtures have to be specific, sensitive, and repeatable, allowing the detection of molecular fingerprints to identify and quantify chemical species. Among them, the NMR spectroscopy is a powerful tool to analyse complex mixtures because structural information and quantification, with its linear response, are possible to extract within one spectrum. Generally, mixture analysis has two main objectives: the

identification and the quantification of the components present in a mixture. With NMR, the identification is performed by comparing the position of NMR peaks in ppm with a reference database. The quantification is purchased with an internal or external standard taking into account the detection limit of the NMR of the order of µM to a hundred of nM (at 600 MHz with a standard inverse probe).

The mixtures can be more easily analysed by hyphenated methods using the help of other techniques where the samples can be physically separated with chromatography (*e.g.* LC-NMR),[59] or analysed with sophisticated NMR methods, such as multidimensional experiments, to overcome the lack of resolution in 1D spectra. The developments of this technique have led to greatly enhanced resolution using experimental techniques such as COSY, TOCSY, HSQC, HMQC, and HMBC which are commonly used for complex mixtures. More sophisticated experiments, such as 3D HSQC-TOCSY,[60] 2D DOSY,[49] 3D DOSY,[61] and Maximum-Quantum (MaxQ) NMR,[62][63] are used to obtain different information. Broadband homodecoupling techniques, also known as pure shift NMR methods, aim to disentangle overlapped spectra by suppressing the effects of homonuclear coupling interactions to turn multiplet signals into singlets.[64][65][66] Unfortunately, the costs in sensitivity and experiment time can be high and increasing the spectral resolution, the freedom from signal overlap is not guaranteed at all. Molecule mixtures can also be processed with different decomposition techniques which, however, are not commonly used due to software difficulties, not simple literature explications, and lack of examples.

The growing complexity of the samples leads to crowded spectra that compromise the analytical performances of NMR. Despite all the methodological developments of NMR, the analysis is sometimes still very challenging, especially for signal assignment.[67] More precisely, this thesis is focused on mixtures of small molecules (less than 1kDa) by implementing different approaches by NMR.

I.1.3.1. Mixture analysis by DOSY

DOSY spectra are an important approach for mixture analysis by NMR.[49] As explained in paragraph I.1.2.4., in this experiment, the peaks belonging to the same compound should decay in the same way. The problem comes when many overlaps are present and it cannot be entirely seen how the diffusion should be.[67] In this situation, multivariate processing is useful and the covariance helps to see what happens in diffusion experiments.

It can be supposed that DOSY data compose an X-matrix, that is a table in which each row is a decaying spectrum. This data set can be described as a product of the decays (C-matrix) and the individual spectra (S-matrix): $X = CS^T + E$, where X is the mixture matrix, C the mixing matrix, S the matrix of the pure spectra, ^{T} denotes the transpose, and E is the residual matrix with the experimental error, which is the noise (**Figure 22**).

Figure 22 *Schematic illustration of the decomposition of an experimental DOSY data set X into component spectra S and decays C. Reprinted with permission from Nilsson et al.[68] Copyright 2008 American Chemical Society.*

In the matrix model, the mixing matrix C contains the associated set of diffusional decay shapes as a function of gradient amplitude. It is easy to simulate the X-matrix, from the product CS, but it is more difficult to find C and S from the X data set.[69]

A typical way of handling multivariate data is with Principal Component Analysis (PCA).[70] The good strategy is to impose constraints. One of them is the imposition that the decay is going to be exponential.

Two tools for processing DOSY data, the *DOSY Toolbox* and the *General NMR Analysis Toolbox* (GNAT) were developed by M. Nilsson's group that include algorithms like DECRA, MCR, OUTSCORE, SCORE, PARAFAC.[71,72] PALM-software also uses an improved algorithm for DOSY signal processing developed by M. A. Delsuc's group available online.[73] In most DOSY spectra, a lot of overlaps are present, which are confusing and difficult to interpret, thus different approaches are possible. Among them, the SCORE method uses the minimization of residuals in matrix $E=X-CS^T$ assuming a known decay form.[68] This is the most basic and one of the most efficient approaches if the diffusion coefficients are not so close and the overlaps are limited. The OUTSCORE method (Optimized Unmixing of True Spectra for COmponent REsolution) is used to un-mixing DOSY spectra with overlapped peaks for similar species, improving the situation by almost an order of magnitude.[74] OUTSCORE analysis provides minimization of the spectral similarities, using the product $E = |Si|.|Sj|$. It acts by combining least squares, fitting cross-talk minimisation, and maximising spectral difference. SCORE needs more than 20% of difference in diffusion coefficients when the overlaps are not so strong. By minimizing the cross-talk in OUTSCORE, until 5% of difference in diffusion coefficients can be resolved.[75]

Another multivariate analysis is Direct Exponential Curve Resolution Algorithm (DECRA)[76,77] which is a fast method to calculate decay rates for each component and to extract spectra mathematically, which does not require a threshold. It is less sensitive to noise levels and deals with moderate spectral overlaps. Deviations from the directly measured values (*e.g.* diffusion coefficients) are related to the fact that the algorithm did not completely purify the component spectra.[76] The data need to be of very high quality to avoid problems. In this context, reference deconvolution methods aim to correct the spectral distortion due to the convolution by some instrumental lineshape function *L(w)*. The corrected spectra are obtained by replacing the instrumental function *L(w)* (deconvolution) by an ideal lineshape *I(w)* (reconvolution).[78]

Instead of using multivariate analysis of DOSY data, multiway analysis is another possibility. As mathematical contribute, a diffusion spectrum constitutes a bilinear data set, a table, or a matrix, with diffusion in one dimension and chemical shifts in another. If a 3D data set is available, with one extra dimension, then the mathematical representation will be with another component added to the model (**Figure 23**).
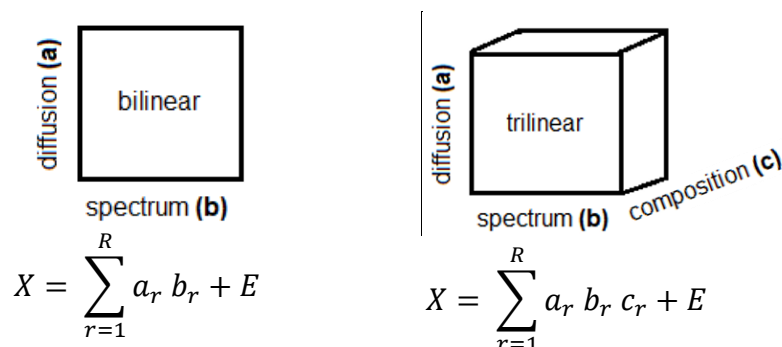

$$X = \sum_{r=1}^{R} a_r \, b_r + E \qquad\qquad X = \sum_{r=1}^{R} a_r \, b_r \, c_r + E$$

Figure 23 *Diffusion spectrum constitutes a bilinear data set. A 3D data set as another extra dimension, thus the mathematical representation has another component. Reprinted with permission from Matthias Nilsson's talk at EUROMAR 2021 conference.*

The difference with the 2D data set is that for trilinear data set it does not need to add any constraint. Parallel Factor Analysis (PARAFAC)[79,80] uses an additional independent dimension to obtain trilinear data by varying three dimensions independently. The third component could be related by the difference in concentrations within recorded samples, by a natural or synthetic way,[81] during a chemical reaction,[82,83] or change in concentrations from sample to sample. PARAFAC requires only knowledge on the number of components to fit.

Another general alternative to process PFGSE data is the maximum entropy reconstruction (MaxEnt).[84] This is an approach that relies on finding the most probable Laplace spectrum by maximizing the entropy of the spectral distribution subject to certain constraints.

Overall, all these decomposition approaches have some problems in estimating sources, and most of the time they require prior knowledge such as the number of sources.

I.2. Signal Processing and Mathematical Decomposition

I.2.1. The Blind Source Separation paradigm

Despite the methodological developments and the simplification of spectra, when NMR is challenged with very complex mixtures, it very quickly reaches its limits in terms of identification and quantification. One-dimensional spectra become unusable and the interpretation of two-dimensional NMR spectra is often very difficult and time-consuming or even impossible. This is why we choose to apply mathematical tools coupled to the NMR to decompose mixtures into source spectra.

The aim of Blind Source Separation (BSS) is to use data processing methods to recover a set of pure signals, called sources, which constitute the spectra of the pure molecules, and their concentrations, starting from a set of linear mixtures. This is performed with limited information or without a prior knowledge of the system, especially regarding the sources and the mixing process, hence in a "blind" way.

The BSS is an inverse problem where we are faced up with the process of calculating, from a set of observations (the mixtures of departure), the factors that produced them (which are the components and their concentrations).[1] It is called "inverse" because it starts with the effects and then calculates the causes. Inverse problems are some of the most important mathematical topics because they tell us about parameters that we cannot directly observe. They have wide application, and especially the blind separation was originally used for disciplines such as telecommunications with Cardoso,[2] and neurosciences with Jutten and Herault in the 1990s.[3] Other applications were carried out in acoustics for source reconstruction,[85] for speech signals or bird chirpings, for example,[86] and in many other fields, such as image processing.[87]

In this approach, the X-matrix, containing the NMR spectra of the mixtures, is modeled as a linear combination of pure spectra, collected in a matrix S, the coefficients, being the

concentrations encoded in a matrix A, with an additional matrix term N, encoding measurement noise (**Figure 24**). Mathematically, this takes the form of a matrix product of these two matrices, complemented by the noise matrix. In the BSS literature, S is called the source matrix, and A is the mixing matrix.
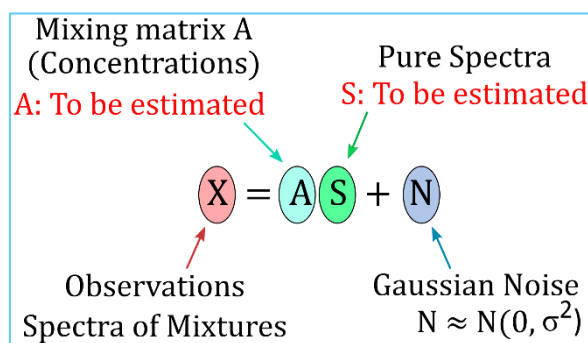

Figure 24 *Mathematical tools: modelling the NMR spectra with the Blind Source Separation algorithms.*

For the decomposition of the spectra, the algorithms use correlations, essentially variations in the concentrations of the molecules through the mixtures. From one sample to another, the intensities or the integrals of the peaks belonging to the same molecule will present the same evolution. In other words, the prerequisite for this extraction is that the samples are recorded in a manner of which the components of interest are purposely varied, here in concentration, either by design or by natural variation. These correlated variations will be detected by the algorithm to extract the pure source spectra, even if there are overlapped peaks. One of the potential applications of this approach will be in metabolomics where a series of samples are available with up to fifty molecules with varying concentrations (**Figure 25**).
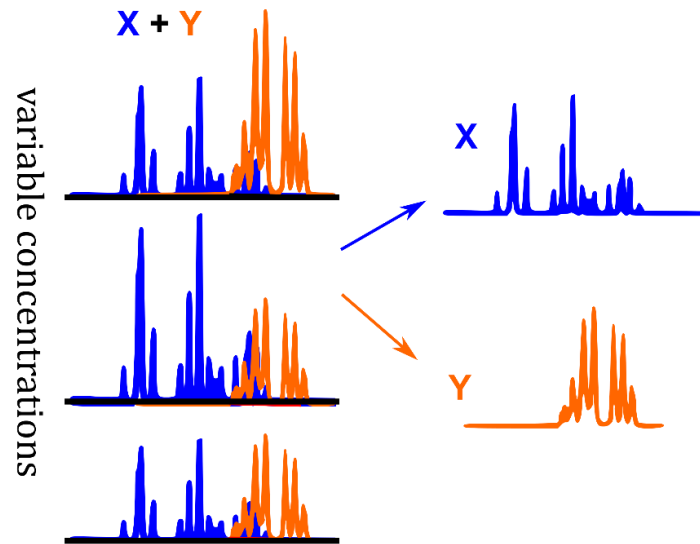
Figure 25 *A scheme to show the mechanism of the BSS. On the left, three mixtures containing two pure sources, the orange and the blue, of different concentrations or peaks integrals. In the spectra there are overlapped parts and isolated regions. On the right, each spectra is extracted for each estimated sources.*

I.2.1.1 Problem statement

The goal of the BSS in NMR is to identify the composition of the mixtures, by extracting the spectrum of each compound. For this purpose, we assume that the observations, composing the matrix *X* of mixtures, are related to the matrices *S* and *A* (**Equation 1.8**).

$$\boldsymbol{X} = \boldsymbol{AS} + \boldsymbol{N} \tag{1.8}$$

where all quantities are represented by matrices, and *AS* is the standard matrix product. More precisely:

- $X = (x_{m\ell}) \in \mathbb{R}^{M \times L}$ is the matrix of observed mixtures: its rows are the *M* measured signals and *L* is the number of processed points in acquired spectra (chemical shifts, expressed in ppm or Hz, composing the length of the spectra)

- $S = (s_{k\ell}) \in \mathbb{R}^{K \times L}$ is the matrix of unknown sources, whose rows are the *K* spectra of compounds

- $A = (a_{mk}) \in \mathbb{R}^{MxK}$ is the mixing matrix, whose M rows label mixtures and L columns label compounds

- $N = (n_{m\ell}) \in \mathbb{R}^{MxL}$ is the matrix representing acquisition noise, which has the same dimension as X.

This model (**Equation 1.8**) is in the form of the so-called Linear Instantaneous Mixture (LIM) model. Even though departures from LIM are possible (for example, non-linear effects may occur in several places in the acquisition process), this model is generally accurate enough, and we will limit to it this work.

Element wise, the model in **Equation 1.8** may also be written as:

$$X_{m\ell} = \sum_{k=1}^{K} a_{mk} s_{k\ell} + n_{m\ell} , \qquad (1.9)$$

$$\text{with:} \quad m = 1, \dots M, \quad \ell = 1, \dots L$$

Under the name of the BSS, one can include a large variety of algorithms adapted to various domains and various possible assumptions on A and/or S. The simplest approach in the BSS, in which we are interested here, is to consider our problem as a LIM model, where the observed mixtures are linear combinations of the sources. At this point, we are faced with two possibilities: simulated and real cases. In the so-called instantaneous BSS problems, no extra transformation is performed on the sources prior to mixing and creating simulated mixtures. The simulated data are mixtures mathematically generated from known matrices S and A. This leads to the generation of a non-noisy observable X as a function of A, S, under a well-defined model:

$$X_{sim} = AS$$

The generation of a simulated noisy X is also possible, as a function of A, S if we consider the mean Gaussian noise equal to zero and corresponding to σ, the standard deviation of the noise, chosen in advance, where:

$$X_{sim} = AS + N$$

In the real case, this applied model is not well-defined but is only approximate. In a real situation we have to be able to extract S and A from a given observation matrix X. There are more factors that can interfere such as spectrometer and experiment factors (nc_proc, rg, *etc.*), hence we use an observation model which is only approximate in a real situation:

$$X \approx AS + N$$

Pre-processing, in particular baseline correction, make the assumption possible that the noise is zero mean, and justifies the approximation:

$$X \approx AS$$

The BSS problem, with various formulations, aims mathematically to estimate jointly the source matrix S and the mixing matrix A from the sole observation matrix X.

A BSS problem is overdetermined if the number of observations m is greater than the number of the sources k to estimate (m mixtures $> k$ sources), determined if $m = k$, otherwise the problem is called undetermined ($m < k$).[57]

Different assumptions or models can lead to different identification algorithms, among them, we may mention statistics-based approaches such as Independent Component Analysis (ICA) and Second-Order Blind Identification (SOBI), or Non-negative Matrix Factorization (NMF).

I.2.2. State of The Art for Blind Signal Decomposition

Blind signal processing has received attention during the last thirty years by C. Jutten,[3] J. F. Cardoso,[2] A. Belouchrani,[88,89] P. Comon,[90,91] and others[92–101] due to its wide range of potential applications in science and mathematics.[57,92] These signal decomposition methods are already very useful in a number of fields such as speech processing or acoustics in general, seismic exploration, digital communications, mobile communications, radar, and in military applications such as interception and classification of radiating sources.[92,90,93] Mobile communications need channel identification which is performed with the help of known sequences, thus involving semi-blind methods.[90] BSS was initially developed in the context of multiple sensor signal processing, in which a set of sensors or channels (observations) receives signals from sources, with intensities depending on their relative positions. Since the data analysis produces the source signals and the mixing coefficients, different assumption or models can lead to different identification algorithms.

Among them, statistics-based approaches such as SOBI (Second Order Blind Identification, which exploits second-order moments) or the many variants of ICA (Independent Component Analysis, which exploits higher-order statistics)[57,90,102] are probably the most widely used.[103] In these approaches, one searches for an "unmixing matrix" that optimizes some statistical criterion (uncorrelatedness or independence of estimated sources).

The recent developments in numerical optimization have motivated the use of variational approaches, in which the problem is formulated as the problem of minimizing with respect to A and S a suitable objective function. The latter often involves constraints or regularization, and sparsity enforcing constraints or regularizations have been proven efficient in many contexts. Sparsity essentially means that the solution is characterized by a small number of parameters in a suitably chosen representation (**Annexe.1**).[57]

Belouchrani *et al.* in 1997 exploited the time coherence of the source signals with the algorithm based on second-order statistics (SOBI), designed to deal with temporally correlated signals.[2] Their algorithm was used for the application of the BSS on NMR spectra by Nuzillard and co-workers published in 1998.[2, 104] In this work, applications on 1D ^{13}C resolved spectra and 2D HSQC NMR spectra were performed. 2D results were not satisfying and the quantitative aspect was not assessed, however, it was the inspiring work to start our project from I. Toumi until us.[57]

Before us, the application of the BSS on NMR data was carried out in our team on 1D DOSY spectra to improve the processing of these spectra with overlapping species, distinguishing molecules with similar spectra and diffusion constants.[105, 106]

Our work inspired also, last year, R. J. McCArty *et al.,* which published a paper on the BSS for NMR spectra with negative intensity.[107]

Outside the BSS, unknown compound identification can be performed by statistics according to the variations along a series of spectra such as changes in concentrations of components in the mixtures. The peak intensity constitutes an additional dimension thanks to which the spectra are partially decomposed in the Statistical TOtal Correlation SpectroscopY (STOCSY).[108,109]

The use of typical chemometrics tools applied to NMR spectra is another interesting approach.[110] Among all, there is the processing method towards curve resolution, under the name of Multivariate Curve Resolution (MCR), which is most frequently used in diffusion applications. It aims to recover pure decay profiles from DOSY NMR data extracting pure components, but also to de-noise low sensitivity 2D ^{29}Si-Solid-State MAS NMR data.[111,112,113,114] No variations in concentrations were present but only in signal intensities because of the indirect evolution, since the spectra were acquired by varying some experimental parameters in the pulse sequence. This was applied on one sample acquired varying the efficiency of ^{1}H-^{29}Si transfer as function of mixing time, hence varying the cross-polarization

contact-time across different experiments. For MCR, a threshold value has to be added, and the number of components has also to be set for the initialization step.[114]

I.2.2.1. Indeterminacies

It is supposed that the estimated X is a function of S and A. The multiplicity of solutions A and S is computed at each iteration. Some difficulties are present in classical source separation in general, such as the non-unicity or indeterminacy of the solution. In this case, there are also problems related to the high dimension of NMR data and the poor resolution of one-dimensional NMR spectra.

The solution of this problem is not unique, as for any solution (A, S) one can write $X = AS = A'S'$, where $A' = A\Lambda$ and $S' = \Lambda^{-1}S$, for some invertible matrix Λ. There exist infinitely many such invertible matrices, thus infinitely many solutions. Additional assumptions or constraints are then, necessary to restore uniqueness (up to trivial indeterminacies to be discussed below). Among these indeterminacies, two types of matrices Λ can be presented (which represent the above-mentioned trivial indeterminacies).

- Scale indeterminacy or renormalization (Λ diagonal): the sources can only be identified up to a normalization or constant factor. Multiplying a row of $S_{k,l}$ by a non-zero value and dividing the corresponding column of $A_{m,k}$ by the same constant, do not modify $X_{m,l}$, or the product AS.

- Order indeterminacy (Λ a permutation matrix): estimated sources are not ordered. Exchanging two rows of S and the corresponding two columns of A does not change X. The comparison of estimated sources with reference sources has to be preceded by an ordering step.

I.2.2.2. Optimization using Non-negative Matrix Factorization (NMF)

Optimization of the model with additional assumptions is necessary to solve the problem of indeterminacies with respect to the two variables S and A and to stabilize the solution to guarantee the unicity and the convexity.

With Non-negative Matrix Factorization (NMF), source coefficients $s_{k,\ell}$, representing spectrum values, and mixing matrix coefficients $a_{m,k}$, representing concentrations, are non-negative. The objective function is introduced in **Equation 1.10**, which reports the most classical version, where there is the sum of squares (the squared ℓ_2 norm) of the difference $X - AS$, between data in the given X-matrix and the ones derived by the LIM model $X = AS$:

$$F(X|A,S) = \frac{1}{2} \, || \, X - AS \, || \, _F^2 \tag{1.10}$$

Previous approaches to NMF were proposed especially by Lee and Seung,[115] where various formulations with numerical algorithms can be used to minimize the objective function depending on the two variables A and S:

$$\min_{A,S} F(X|A,S), \text{ under constraints } A \geq 0, S \geq 0 \tag{1.11}$$

wherewith NMF all the matrix elements of A and S, and X of consequence, has to be non-negative.

The objective function involves a data fidelity term, linked to the linear model, which forces the product AS to be as much as possible close to the given data of observations in X-matrix. Additional terms, encoding prior information on A and/or S, can possibly be added (for example, limiting the concentration range). The classical version of the objective function is to use penalization terms in the standard quadratic objective function (**Equation 1.12**):

$$F(X|A,S) = \frac{1}{2} \, || \, X - AS \, || \, _F^2 \; + \; f_A(A) + f_S(S) \tag{1.12}$$

where the first term is the squared ℓ_2 norm that uses the least squares, which is the sum of squares of *X-AS*, and f_A, f_S are regularization terms, which encode prior information on *A* and/or *S* matrices.

I.2.2.3. Alternate Optimization

The described non-negativity assumptions made in NMF approaches turn out to resolve a part of these problems, nevertheless, scale and order indeterminacies remain.

The leading strategy is to estimate jointly *S* and *A* as outputs to minimize the objective function noted *F(A, S)*, with respect to *A* and *S* alternatively until convergence of the objective function to a minimum. The objective function is generally chosen as the sum of the data fidelity term and regularization term (**Equation 1.13**).

Regularized approach:

$$\text{minimize}_{A \geq 0,\, S \geq 0}\ F(X|A,S) = \underbrace{F(A,S)}_{\text{Data fidelity term}} + \underbrace{f(A,S)}_{\text{Regularization term}} \qquad (1.13)$$

This problem is non-convex, more precisely bi-convex (*i.e.* convex with respect to *A* and *S* independently). The non-convexity is illustrated with the red line in **Figure 26**, where the objective function *F* is a function of the number of iterations, related to *X* estimated from $A^i S^i$, to reach the convergence as the optimal solution (blue line).
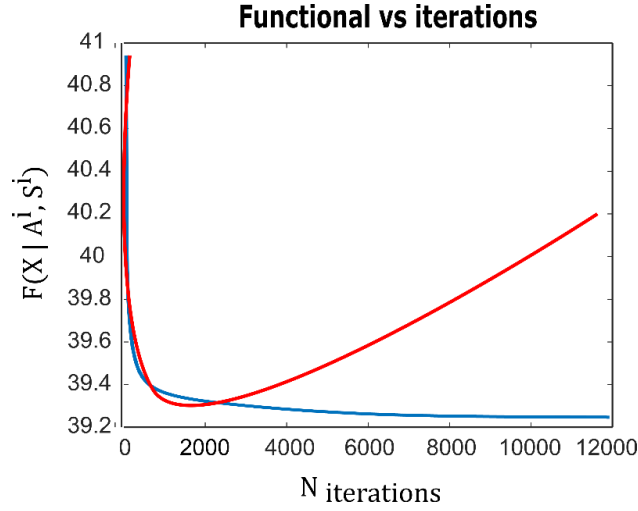
Figure 26 Objective function and number of iterations for the optimal solution with convergence in blue and a stabilized solution, and with the non-ideal case without convergence (divergence) in red.

It is proposed to use the iterative optimization methods where the Majorize - Minimization seems to be a very efficient strategy.[116–118] MM algorithm is described in **Annexe.2**.

The goal is to minimize the average reconstruction error between the given X-matrix, the target, and the estimated *X* function of *AS*. This error has to be minimum; this means that we will look for the source and mixing matrices, *S* and *A*, such that the reconstruction error, i.e. the squared norm of the difference *X-AS*, is minimal. This squared norm (*i.e.* $\|x\|^2 = \sum_1^n |x_n|^2$) is a quadratic function of *A* (*i.e.* a second-order polynomial in the matrix coefficients of *A*), and a quadratic function of *S*. Hence its gradient with respect to *A* is a linear function of *A*, and its gradient with respect to *S* is a linear function of *S* too. Finding the minimum with respect to *A* and *S* amounts to find values for *A* and *S* such that these two gradients vanish. This can be achieved through iterative algorithms, which solve iteratively the two problems one after another.

The bi-convexity of the problem ensures the existence of a global minimum for each unknown (*A* or *S*) when the other is kept fixed, but this is not true anymore when the two unknowns are considered jointly (**Figure 27**).
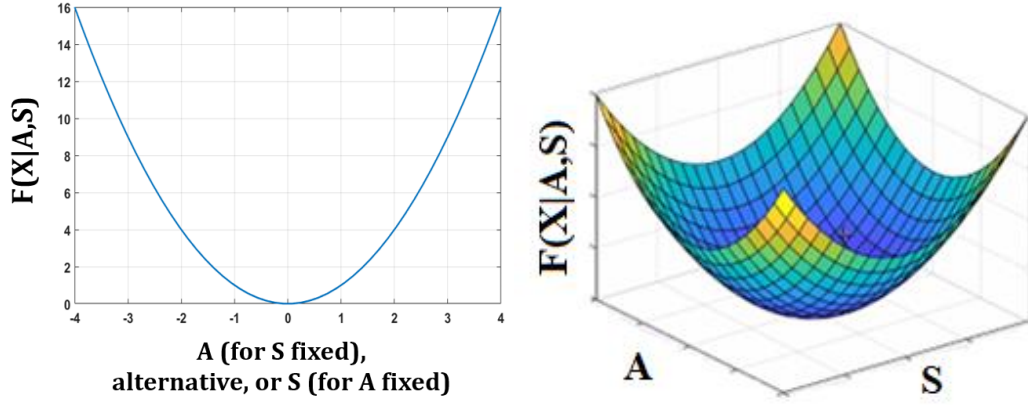
Figure 27 *F(A,S) is function of A and S estimated at each iteration, which may be considered with two functions, once estimating A, once estimating S alternatively (left). If one variable is fixed, the objective function is simply convex. If both A and S has to be found, the problem is bi-convex. Hypothetically this may be a 3D plot, with two functions in the X^2 (right), in case of an optimal solution with a global and unique minimum.*

There may be local minima, or even highly degenerate minima (minima that are reached for infinitely many pairs (A, S) (**Figure 28**).
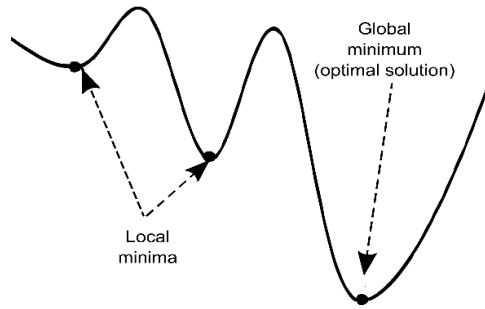
Figure 28 *Local minima coexisting with the absolute minimum in a non-ideal case.*

The situation can be improved when the problem is modified, to involve additional terms in the objective function, whose role is to promote some specific behaviour of the solution. Examples are given by non-negativity constraints (all matrix coefficients of A and S have to be non-negative) and sparsity enforcing penalizations, which promote matrices with few non-zero entries. These are the choices we have made in this work, enforcing non-negativity on both A and S and sparsity on S (**Annexe.1**). The positivity constraint is imposed for the mixing operator A, moreover, we can also strongly constraint A, not only between 0 and $+\infty$, but we can fix an

interval if we want. For the source matrix, we can consider that the exploited data, representing NMR spectra, are always sparse and positive.

To this end, we use the new objective function (**Equation 1.14**) with appropriate penalty functions f_A and f_S:

- f_A is the indicator function of non-negative matrices, equal to 0 for non-negative matrices and +infinity as soon as A has a negative coefficient,

- f_S is the sum of the indicator function of non-negative matrices and a constant λ_S times the ℓ_1 norm. The latter turns out to promote sparse sources, the parameter λ_S controlling the degree of sparsity. In the algorithm, λ_S can be seen as a threshold, the entries of S smaller than λ_S being set to zero.

The problem in **Equation 1.14** involves the classical quadratic data fidelity term, and linked to the linear model, with the regularization functions just introduced, on A and S, represented by f_A and f_S (**Equation 1.14**):

$$\text{minimize}_{A \geq 0,\, S \geq 0}\; \frac{1}{2} \lVert X - AS \rVert_F^2 + \lambda_A\, f_A(A) + \lambda_S\, f_S(S) \qquad (1.14)$$

with λ_A and λ_S, the regularization parameters respectively on A and S (detailed in **Annexe.4**). In order to solve the bi-convexity of the problem, the alternate optimization on A and S was the adopted schema to solve this issue, which cannot be solved as a normal optimization.

I.2.2.4. BSS Generic Algorithm

The algorithms change depending on the choice of the objective function, optimization strategy and source representation domain (spectral or wavelet domain (**Annexes.4-8**)).

The general structure of the algorithm is reported and the rules to optimize A and S change with updates (noted as Upd_A and Upd_S) in the generic form of:

Upd_A : means that just A will be updated $(A, S) \rightarrow \text{Upd}_A (A, S) \in \mathbb{R}^{MxK}$

where M is always the number of mixtures and K the number of the sources.

Upd_S : means that just S will be modified $(A, S) \rightarrow \text{Upd}_S (A, S) \in \mathbb{R}^{KxL}$

where K is the number of sources and L the observations, hence the values at frequency ℓ, composing the length of the spectrum.

 With these updates, the generic structure of the alternate optimization algorithm for NMF is described below and reported in **Figure 29**.

After entering the X-matrix and the number of iterations to arrive at the convergence, the alternate optimization algorithm starts to compute through different steps. The first stage is the initialization with the matrices A_0 and S_0 to be entered. Initial matrices S_{init} and A_{init} are necessary to start the iterations. Some algorithms require only one of these. The BSS algorithms require an initial estimate. Here we used Joint Approximate Diagonalization of Eigen-matrices (JADE) Independent Component Analysis (ICA) as initialization step to obtain estimates used also on the work of I. Toumi *et al.*[57] JADE only requires the number of sources k to be estimated to start the initialization as a function of matrix X. JADE, using correlations and run on X-matrix, produces an estimated output, whose absolute value is used for the un-mixing matrix in order to obtain an estimate for the sources and the matrix A (JADE description in **Annexe.3**).

The algorithm stops when a prescribed maximal number of iterations is reached, or preferably when some precision criterion reaches a small enough value. A possible choice used for the stopping criterion is with the normalized norms of differences between two consecutive iterates of A and S, that is with the objective function update from one iteration to the next:

$$\text{Crit}(k) = \left| \frac{F(X|A^{(j+1)}, S^{(j+1)}) - F(X|A^{(j)}, S^{(j)})}{F(X|A^{(j)}, S^{(j)})} \right|$$

where $A^{(j)}$, $S^{(j)}$ are the estimated matrices at iteration j, and $A^{(j+1)}$, $S^{(j+1)}$ are the ones to the next iteration.

Before submitting the data to the algorithms, some formatting is necessary, for example, the normalization of the data to take into account the experimental variations, and the optional pre-processing to project the data of matrix X to set all negative matrix elements $x_{n\ell}$ to zero: $X(X<0)=0$.

After this description of the generic structure of alternate optimization using NMF, the general schema of the algorithm is reported below in **Figure 29**.

Data : *X*-matrix

- *iter_max*: The number of iterations

Initial data and parameters provided by the user

- A_{init} : initial mixing matrix

- S_{init} : initial source matrix

- ε : the target value for the stopping criterion noted *Crit* (defined below) to end the optimization

- **Stopping criterion** of departure = +∞ (starting from a big value).

Initialization: $A_0 = A_{init}$, $S_0 = S_{init}$; for $n_{iter.} = 0$

Update A: $A^{(j+1)} = Upd_A (A^{(j)}, S^{(j)})$

While: j ≤ iter_max and Crit>ε

Update s: $S^{(j+1)} = Upd_S (A^{(j+1)}, S^{(j)})$

Optimization: normalization of A and/or S;

Evaluation of stopping criterion Crit(j+1)

Evaluation of the objective function $F(X|A^{(j+1)}, S^{(j+1)})$

Crit > ε

Crit ≤ ε

j=j+1

If iter_max

end

Result: Non-negative matrix factors A and S

Figure 29 *Generic structure of the alternate optimization algorithm for non-negative matrix factorization, where this generic algorithm requires additional options.*

I.2.2.5. Why NMR data can be subjected to BSS

To understand how BSS works in NMR, it is important to understand that the NMR spectrum of a molecule can be modeled as a set of chemical shifts δ, with intensities that depend on the number of isotopes that represent each multiplet, and the overall intensity will depend on the concentration of the molecule noted *A*.

$$NMR\ Spectrum = a\ (\ I\ \delta i, J\ \delta j, K\ \delta k, \dots)$$

where: $I\ \delta_i, J\ \delta_j, K\ \delta_k, \dots = S$, correspond to a table like that one reported in **Table 1**.

Table 1 *Illustration of an NMR 1D 1H spectrum represented in tabular form. In this example, the spectrum was divided into 0.001 ppm-wide buckets.*

δ (ppm)	9.9995	9.9985	9.9975	9.9965	9.9955	9.9945	.....
Intensity	40158.5	114354.56	242205.87	390999.75	104130.31	26666.62	.....

The NMR spectrum (in **Figure 30**) can therefore be represented by the product *aS*:

$$NMR\ Spectrum = aS$$
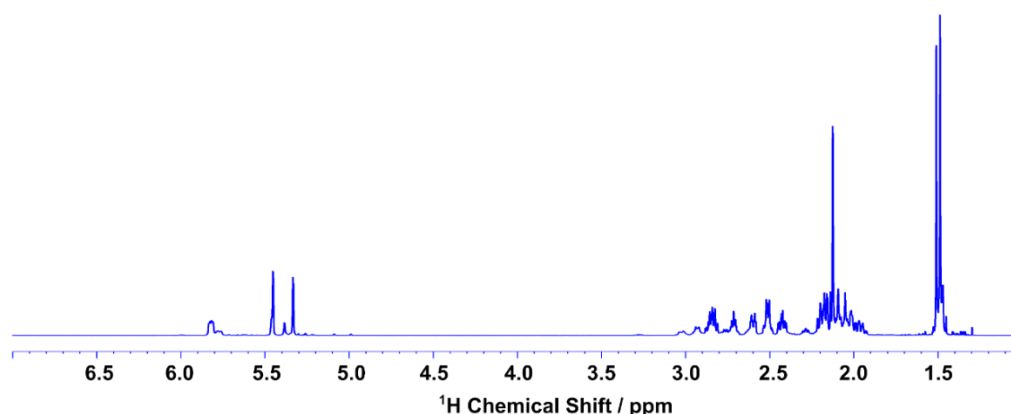
Figure 30 *Example of an NMR spectrum of a molecule.*

To generate noisy mixtures, we add Gaussian white noise, with zero mean and standard deviation σ equal to the standard deviation of the true mixture, estimated from a signal-free region.

In the noiseless case, the spectrum of a mixture of two molecules is the sum of the two NMR spectra of each molecule. This property is related in mathematics to the LIM model, cited above, where the observed mixtures are linear combinations of the sources.

$$X = (a_1 S_1) + (a_2 S_2)$$

In the more realistic noisy case, these properties remain approximately true, as long as the noise is not too large, regardless of the number of molecules in the mixture. An NMR spectrum of a mixture with four compounds, for example, can be modelled as the sum (a linear combination) of the pure spectra of each molecule multiplied by its concentration (**Figure 31**).

$$X = (a_1 S_1) + (a_2 S_2) + (a_3 S_3) + (a_4 S_4)$$
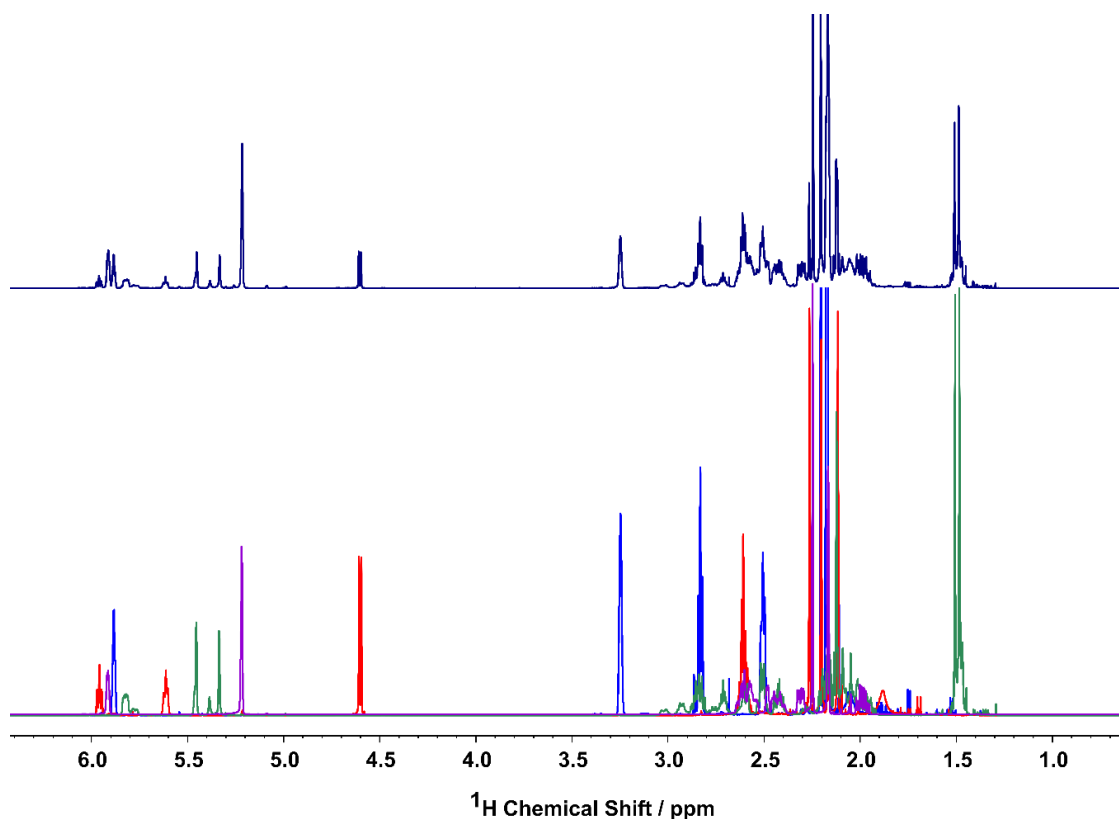
Figure 31 Superposition of the spectra of four molecules (on the bottom), and the mixture that can be obtained by the sum of the four spectra (on the top).

A dataset of several X_m mixtures with the same composition, in which the concentrations of each molecule change from one mixture to another can be represented in a matrix form, as a product of the matrices A and S (**Figure 32**).

$$X_1 = (A_{11} \times S_1) + (A_{12} \times S_2) + (A_{13} \times S_3) + (A_{14} \times S_4)$$
$$X_2 = (A_{21} \times S_1) + (A_{22} \times S_2) + (A_{23} \times S_3) + (A_{24} \times S_4)$$
$$X_3 = (A_{31} \times S_1) + (A_{32} \times S_2) + (A_{33} \times S_3) + (A_{34} \times S_4)$$
$$X_4 = (A_{41} \times S_1) + (A_{42} \times S_2) + (A_{43} \times S_3) + (A_{44} \times S_4)$$
$$X_5 = (A_{51} \times S_1) + (A_{52} \times S_2) + (A_{53} \times S_3) + (A_{54} \times S_4)$$

X_1 = Mixture 1
X_2 = Mixture 2
X_3 = Mixture 3
X_4 = Mixture 4
X_5 = Mixture 5

$$
\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix}
=
\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \\ A_{51} & A_{52} & A_{53} & A_{54} \end{bmatrix}
\times
\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}
$$

Concentrations   Pure spectra

^{1}H Chemical Shift / ppm

Figure 32 *Matrix model of the BSS problem with a data set of five synthetic mixtures of four sources.*

In the 2D case, the model is also represented by the *X*-matrix with the five different mixture spectra, matrix *X* is equal to the matrix *A* with the concentrations of each source within the five mixtures, multiplied by matrix *S* with the spectra of the four sources. For this, 2D spectra have to be reshaped as 1D series, by concatenation of rows (see **Figure 33**).

$$
\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix}
=
\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \\ A_{51} & A_{52} & A_{53} & A_{54} \end{bmatrix}
\times
\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}
$$

X [m, l]
[5, 1024 x 2048]
=
A [m, k]
[5, 4]
X
S [k, l]
[4, 1024 x 2048]
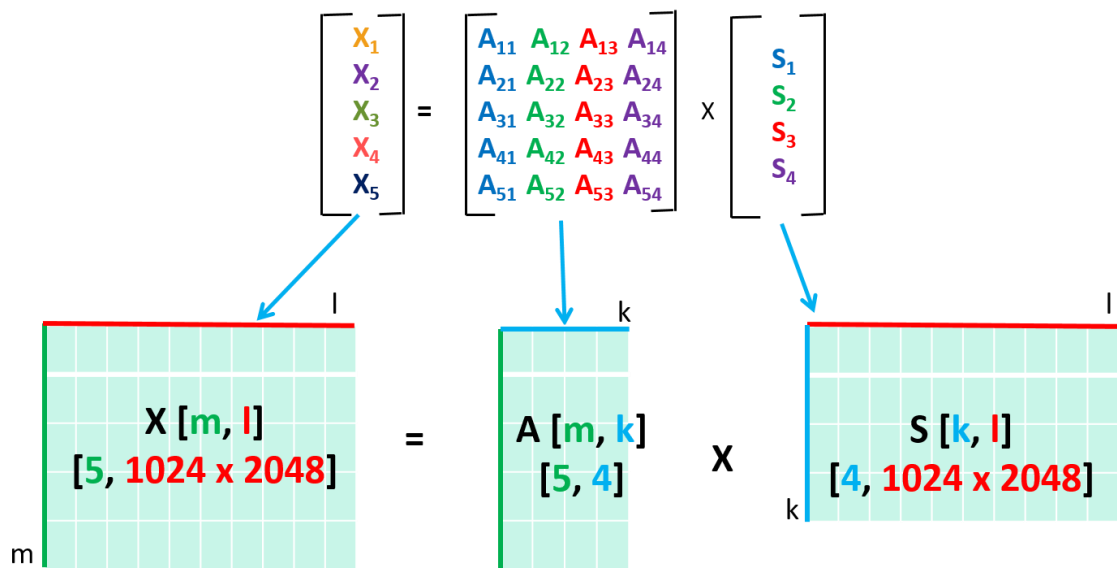
Figure 33 *Schema for the explication of 2D NMR case represented in matrix form.*

I.2.2.6. Dataset Construction

If one gets at least as many linearly independent spectra of mixtures as there are individual components, then it is possible to separate their spectra. Usually, we collect m mixtures $\geq k$ sources to perform the BSS. This is a consequence of the rules followed in matrix multiplication, where the compatibility of the operands (rows and columns of the matrices) is accomplished when the inner dimensions must be equal. The same is for a homogeneous linear system of m equations with k unknowns; if $m \geq k$ hence the unknowns can be estimated. For matrix multiplication, in fact, the number of columns in the first matrix must be equal to the number of rows in the second matrix. The resulting matrix has the number of rows of the first and the number of the columns of the second matrix, as shown in **Figure 34** in 1D case, and in **Figure 35** in 2D case.
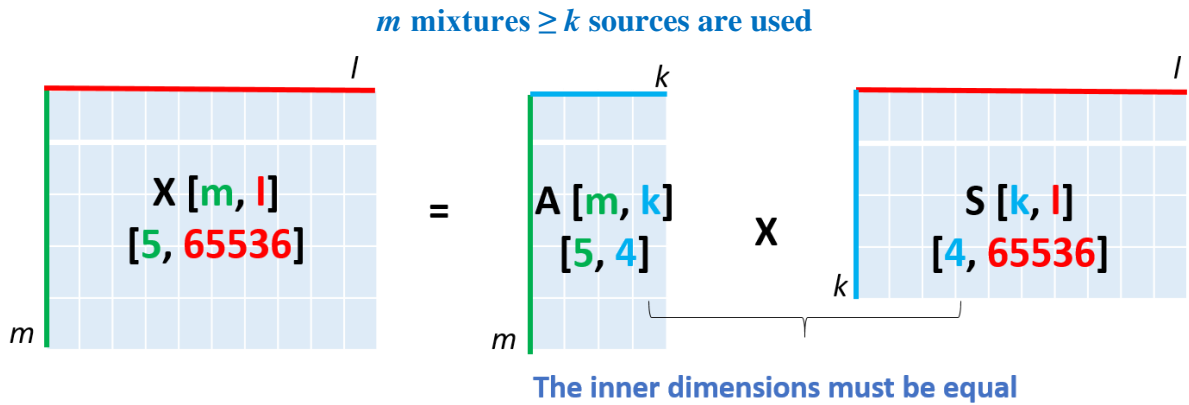
Figure 34 One-dimensional situation illustrating the matrix dimensions. Data matricization is not needed in 1D data sets because the matrices S and X are 2D arrays, hence matrices. 65536 is the size of real spectra in processing parameters, composing the length of the spectra in matrices S and X.

According to the rules in Matrix Multiplication dimensions

Compatibility of operands
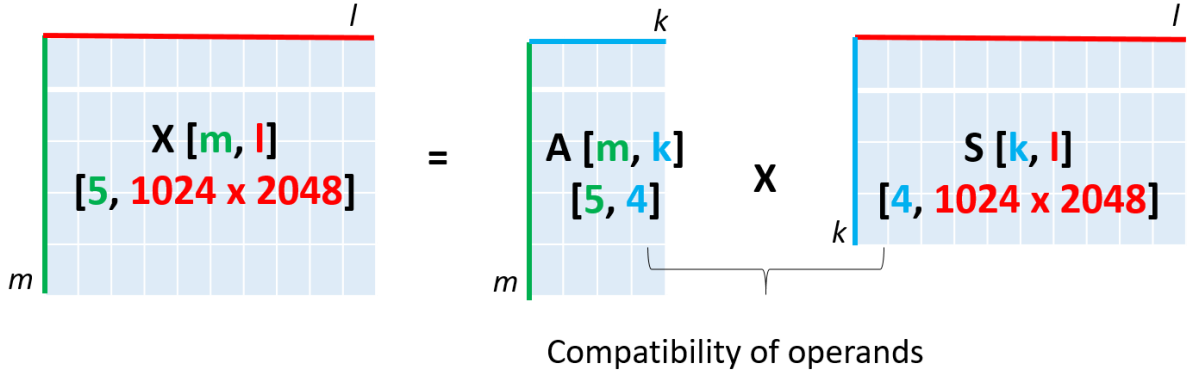
The inner dimensions must be equal :
The shape rule

Figure 35 *In 2D case, the matrices S and X are 3D arrays. They need matrix reshape, hence the matricization step, multiplying 1024 x 2048 which are the size of real spectra in processing parameters (SI in ProcPars), respectively in F_1 and F_2 frequency axes. In matrix multiplication, the operands must be compatible and the inner dimensions must be equal (same shape), in fact, the number of columns in the first matrix must be equal to the number of rows in the second matrix.*

I.2.2.7. Evaluation Criteria

To evaluate the quality of un-mixing results, numerical indexes are used to pursue an objective assessment. The separation algorithms aim to estimate $\hat{A}$ and $\hat{S}$ to be compared with the real S and A in a situation where it is possible to recover or acquire them by NMR.

To assess the quality of estimated sources, the estimation error between real S and estimated $\hat{S}$ is evaluated using two numerical performance criteria with two types of error in S-$\hat{S}$. Among them, Sources to Interferences Ratio (SIR) provides a quantitative evaluation of crossover terms after separation, which are the interferences due to the possible presence of spurious peaks from other sources in the pure spectra. It is used to evaluate if the sources in estimated $\hat{S}$ are well separated and it is defined as:

$$SIR = 10 \log_{10} \frac{||S_{target}||^2}{||e_{interferences}||^2}$$

Sources to Distortion Ratio (SDR) quantifies the level of global distortion due to mixing and separation processes. The SDR is defined as:

$$SDR = 10 \log_{10} \frac{||S_{target}||^2}{||e_{interferences} + e_{noise} + e_{artifacts}||^2}$$

The best estimation corresponds to the highest measurements of these indexes and shows how much S and $\hat{S}$ are closer. They are expressed in decibels (dB) and are graded to a logarithmic scale as for the traditional Signal-to-Noise Ratio (SNR). These measurements are inspired by the usual definition of SNR, with a few modifications, as they express as well the logarithmic ratios of two physical quantities.[85]

The quality of the estimated mixing matrix $\hat{A}$ is evaluated by computing the Amari index which ranges between 0 and 1. When Amari is equal to 0, the best estimation of $\hat{A}$ is obtained, and $A=\hat{A}$. When A and $\hat{A}$ are maximally different, the Amari index is equal to 1.[119] Furthermore, it is important to mention that the Amari index is not sensitive to reordering nor rescaling of sources (the above-mentioned trivial indeterminacies).

II. Evaluation of recent BSS algorithms on 1D and 2D NMR spectra of synthetic mixtures

Blind source separation methods, and more generally blind signal processing, include various approaches which exploit different assumptions. Our approach was introduced and detailed on *Faraday discussion*s on 2019.[5]

The first application of the BSS on NMR spectra was carried out by Nuzillard *et al.* in 1998 and it was applied on the ^{13}C spectra and two-dimensional ^{1}H-^{13}C HSQC spectra.[104] SOBI statistical-based method was used, imposing constraints on the time-shifted source signals, where each source was time-correlated and the noise generated was time-uncorrelated. A choice of the times for the analyses was done with prior knowledge of the time correlation properties of the pure compounds or sources during the reaction. 1D ^{13}C NMR spectra to be separated do not overlap significantly and, furthermore, ^{13}C resonances lines are narrow enough to limit the peak superimposition. 1D spectra dealt with the isomerisation of α-glucose into β-glucose in D_2O where source mixing was considered a linear process, and the number of sources k to estimate was set to two using five m time-domain signals (where $m \geq k$). The second application regarded the separation of 2D HSQC spectra of three mixtures of three sources: sorbitol, mannitol, xylitol in D_2O with high concentrations in the range of 40-60 mM. However, there, only a qualitative aspect was discussed but it was the inspiration to start this project from I. Toumi until this work.[57]

Before us, the application of the BSS on synthetic (artificial) NMR data was carried out in our team by I. Toumi,[105,106] to improve DOSY processing with overlapping species, to distinguish molecules with similar spectra and diffusion constants. Two BSS techniques were presented. The first one was the variational approach, introduced there for the first time, born from the NMF approach using additional Sparse Conditioning (SC), giving the NNSC algorithm. The second one was the Joint Approximate Diagonalization of Eigenmatrices (JADE), declination of ICA, based on statistical modelling. JADE algorithm is computationally faster and simpler to use, and it can be a good tool for DOSY processing of mildly overlapping spectra, but fails

with the increase of the complexity of separation such as overlapping signals and similar diffusion coefficients. NNSC performs well at a high degree of difficulty although with still some imperfections. In **Figure 36** and **Figure 37** are reported the estimated sources from 1D and 2D DOSY spectra using both algorithms. The two methods can be considered complementary as JADE could be used to quickly estimate the number of sources in the mixture in some seconds, to accelerate the NNSC separation. Both methods however fail for too low signal-to-noise ratio levels, but NNSC is more tolerant.
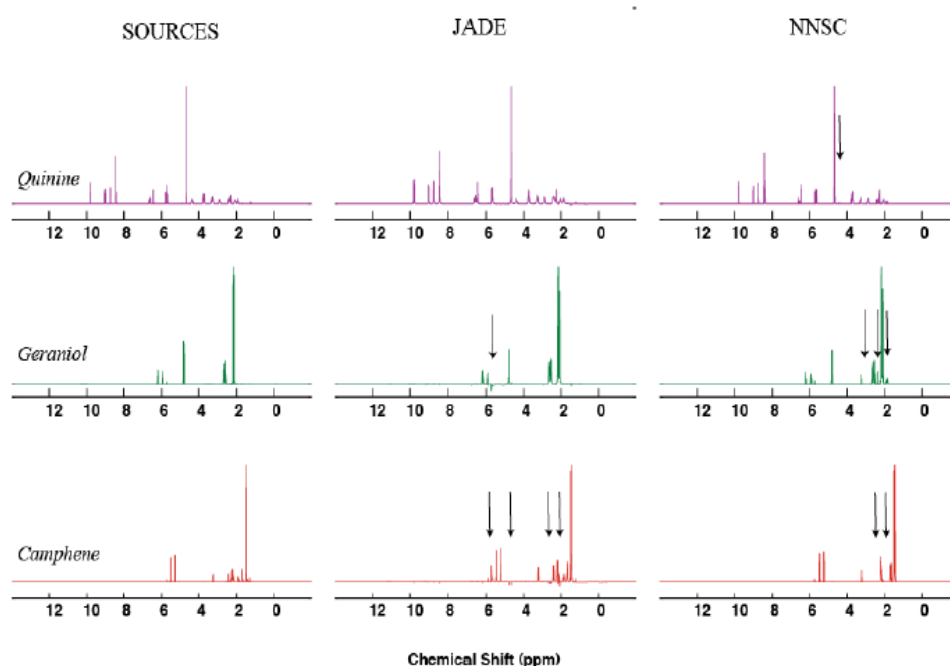
Figure 36 1D ^{1}H PFG-NMR real (on the left) vs estimated sources by JADE and NNSC (on the right). Reprinted with permission from I. Toumi et al.[105] Copyright 2014 Elsevier.
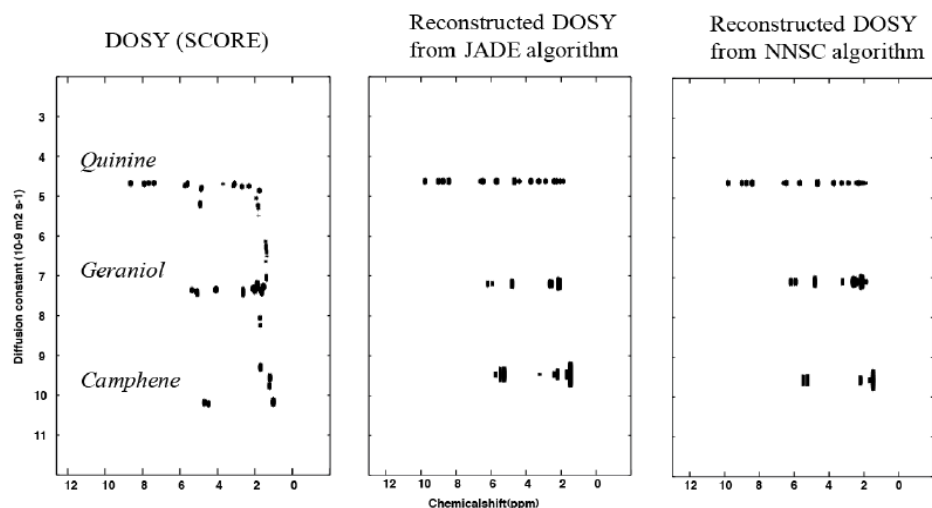
Figure 37 *The reference spectrum (on the left) and the reconstructed DOSY spectra(on the right) of the for QGC mixture from mono-exponential fitting of the peaks in the 2D ^{1}H PGF-NMR experiments, compared with JADE separation and NNSC algorithm. Experiments were carried out on a 400 MHz spectrometer. Reprinted with permission from I. Toumi et al.[105] Copyright 2014 Elsevier.*

The work was improved by introducing a novel sparsity penalized β-NMF (β-SNMF) method, a multiplicative approach, with increased convergence speed by the β-dependent power exponent and more robustness to noise for 2D DOSY spectra.[120]

Another application was performed on fine-needle aspiration biopsies (FNAB) and plasma from thyroid cancer using CPMG relaxation-edited T_2 spectra. JADE and NNSC algorithms can separate broad peaks from sharp ones with some imperfections such as negative peaks.[57] Potential applications of this example can be carried out for complete analysis in metabolomics.

II.1. BSS on Dataset with Concentration Variation

As seen previously, the obvious strategy for blind source separation is to create the equivalent of nD spectra by relying on changes in signal intensity along a series of samples. The most evident come from variations in concentrations, whether natural or artificial. Several algorithms can be used to blindly decompose NMR spectra, among them, projected alternate least squares (PALS), soft threshold projected alternate least squares (STALS), proximal alternating linearized minimization (PALM), block-coordinate variable metric

forward-backward (BC-VMFB) (using spectral domain), and wavelet-based variants on these techniques, where the principals and features are reported in **Annexes 5-9**.

II.1.1. Samples under Investigation: Terpenes

For the BSS applied to 1D and 2D NMR, a mixture data set of terpenes was analysed. Terpenes are natural molecules produced predominantly by plants, presenting highly crowded NMR spectra between 1 and 2.5 ppm, and a characteristic region between 4 and 5.5 ppm. Four commercially available terpenes were purchased from Sigma-Aldrich (Merck) and Saint Quentin Fallavier, France.

- **Terpenes:** (R)-(+)-limonene, nerol, α-terpinolene, β-caryophyllene((-) -trans-caryophyllene) (**Figure 38**).
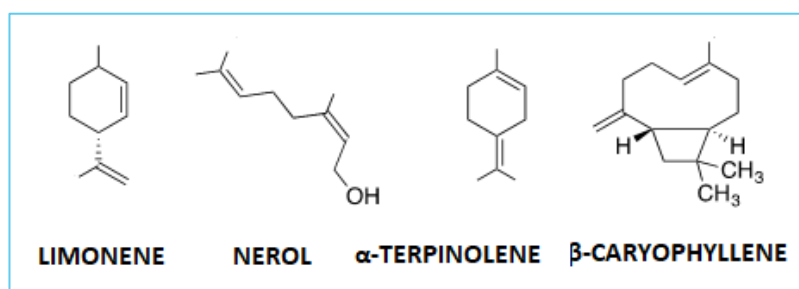
Figure 38 *Chemical structures of terpene molecules used in this study: (R)-(+)-limonene, nerol, α-terpinolene, β-caryophyllene, (-) -trans- caryophyllene.*

The pure compounds were dissolved in 600 μL of $CDCl_3$ at respective concentrations of 181 mM, 36,5 mM, 26,6 mM, and 43,7 mM, and transferred to 5 mm NMR tubes which were sealed to prevent loss of solvent. The samples were therefore stored at -4°C until the NMR characterization. Five synthetic mixtures of the four terpenes were then prepared, varying the concentrations of each compound as reported below with the recalculated ERETIC2 concentrations.

The studied mixtures are considered well adapted to evaluate the source separation algorithms since the highly crowded spectral region is present.

II.1.2. Using Electronic Reference (ERETIC2) to Accurate Determination of the Concentrations

ERETIC2 is a quantification tool with an external standard, which replaces the ERETIC (Electronic Reference To access In vivo Concentration) method. In this thesis, all the concentrations were recalculated using the ERETIC2 method (**Table 2**).[121,122] An NMR spectrum of TSP (TrimethylSilyl Propionic acid) of 7,3 mM in D_2O was collected. The reference signal was integrated and defined as ERETIC2 reference. The concentration of the reference sample was defined (7,3 mM). The number of nuclei for signal (9 H), sample volume (600 µL), and molecular weight (172,27 g/mol) should also be defined.

Table 2 Concentrations of each component of terpenes (in mM) in the five synthetic mixtures in sealed NMR tubes obtained by ERETIC2 method.

	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Mixture 1	23.30	26.00	8.78	10.87
Mixture 2	17.10	11.93	15.50	15.00
Mixture 3	9.05	14.23	18.89	4.67
Mixture 4	20.99	6.86	13.54	11.96
Mixture 5	4.88	9.01	10.81	13.15

II.1.3. nc_proc - Intensity Scaling Factor

Processing in Topspin software performs calculations in double-precision floating point but stores the result in 32-bit integer values. During double to integer conversion, the data are scaled up or down such that the highest intensity of the spectrum does not exceed values that can be supported. This means that the 32-bit resolution is not entirely used. This allows for the highest intensity to be increased, for example during phase correction, without causing data overflow, and it can be used in 1D, 2D, and 3D datasets.

For this, an integer value, only existing as a processing status parameter, is taken. NC_proc shows the amount of scaling that was done. For example, for nc_proc = -3 data were scaled up (multiplied by 2) three times (2^3), for nc_proc = 4 the data were scaled down (divided by 2) four times (2^{-4}). Although nc_proc is calculated automatically, it is possible to known and change its value using the "nc_proc" command. 2D processing also allows to change the scaling factor with the argument "xfb nc_proc 2" for example, scales down the data 2 times. However, one can only scale the data more down than the command would have done without the argument "nc_proc". Smaller (more negative) values of nc_proc are ignored to avoid data overflow.

In conclusion, it is important for 2D spectra to have the same nc_proc value and to scale down to the same higher positive value among all the spectra considered.

Here is an example: for the terpene dataset containing the 2D HSQC US spectra, the highest value of nc_proc was equal to -1 for the limonene HSQC (**Table 3**). We could scale down all the spectra using the command "xfb nc_proc -1" to obtain uniform intensities, but we choose nc_proc= 0 to anticipate eventual down scaling due to an important intensity for the following experiments.

Table 3 *nc-proc values for 2D-HSQC original spectra of sources and mixtures.*

2D HSQC US	Limon.	Nerol	α-terpin.	β-caryoph.	Mix. 1	Mix. 2	Mix. 3	Mix. 4	Mix. 5
nc_proc	-1	-4	-5	-3	-4	-4	-5	-4	-5

II.1.4. Save and Reading of NMR Data Sets

First of all, it is necessary to convert, save and export NMR data from TopSpin (Bruker) to JCAMP-DX format with the suffix '.dx'. The possible output types are FID (FID), real spectrum (RSPEC), and complex spectrum (RSPEC+ISPEC). We used RSPEC to work only

on the real signal. Regarding the compression mode, there are different formats: FIX (table format), PACKED (no spaces between y values), SQUEEZED (sign coded into the first digit of the value, no spaces), DIFF/DUP (coding the differences of subsequent values in the SQUEEZED format with suppression of repeated values). In this work, we always use FIX format.

II.1.5. Data reshaping in matrix form

The matrix A is mxk dimensions that means 5 rows and 4 columns (for the four terpenes), matrix S is kxl dimensions which means that each one of the four rows represents a monodimensional spectrum with its 65536 processed points to reconstruct the real spectra. In the 1D case, matrices X and S take the form of two-way arrays with each 1D spectrum aligned in a line as a vector.

In 2D data sets such as in the HSQC spectra, the observations in X and the pure signals in S are not matrix-shaped and actually give rise to 3D matrices. Source matrix will be $S \in \mathbb{R}^{KxL1xL2}$ where k are the numbers of sources and L its 2D spectrum ($L_1 X L_2$), and $X \in \mathbb{R}^{NxL1xL2}$. L is constituted by 1024 multiplied by 2048 ℓ values constituting the size of real spectra, respectively in F_1 and F_2 frequency axes. One can represent each spectrum in a row as a vector in the matrices S and X. In other words, we have to linearize each source 2D spectra in the same line as a sum of the line blocks composing the 2D spectrum. As we transform all the n-dimensional arrays into matrices, this approach of reshaping is called data matricization.

Here, the LIM model can be re-written as (**Equation 2.1**):

$$X_{ml_1l_2} = \sum_{m=1}^{M} a_{mk} s_{k\ell1\ell2} + n_{m\ell} \tag{2.1}$$

with $m = 1, \dots M, \quad \ell_1 = 1, \dots L_1, \quad \ell_2 = 1, \dots L_2$

where ℓ_1 and ℓ_2 label are the two spectral dimensions. By re-organizing the ℓ_1 and ℓ_2 spectral dimensions into a single one of length L_1L_2 transforming the 3D arrays into matrices, one can be back to the initial LIM model.

II.2. Results on 1D NMR Spectra of Terpenes

II.2.1. 1D Simulated Data: Model Validation

One-dimensional spectra were firstly evaluated in simulated cases to validate the model. In this approach, *X*-matrix is generated from given matrices *S* and *A*, linearly combined according to the LIM model, with a zero-mean Gaussian white noise *N* (according to the model $X_{sim}=AS+N$). The standard deviation σ of this noise was set equal to the one of the experimental noise, estimated from a signal-free segment of the real mixtures. This simulated dataset was reconstructed from the matrix *S* collected from the NMR spectra of the pure terpenes, where the concentrations were calculated for *A* (**Table 2**) and organized in a 5x4 matrix. The four algorithms described in **Annexes.5-9** (PALS, STALS, PALM, and BC-VMFB using wavelets or not) were run on this simulated dataset.

Independently, a Principal Component Analysis (PCA) can be carried out to identify intrinsic patterns within the samples of a data set.[123] Samples with similarities will thus be clustered in the frame of the calculated principal components, *i.e.* linear combinations of the NMR variables. Here, the PCA of the *X*-matrix shows that only four out of five principal components are significant, which suggests setting to four the number of terpene sources to estimate (**Figure 39**).
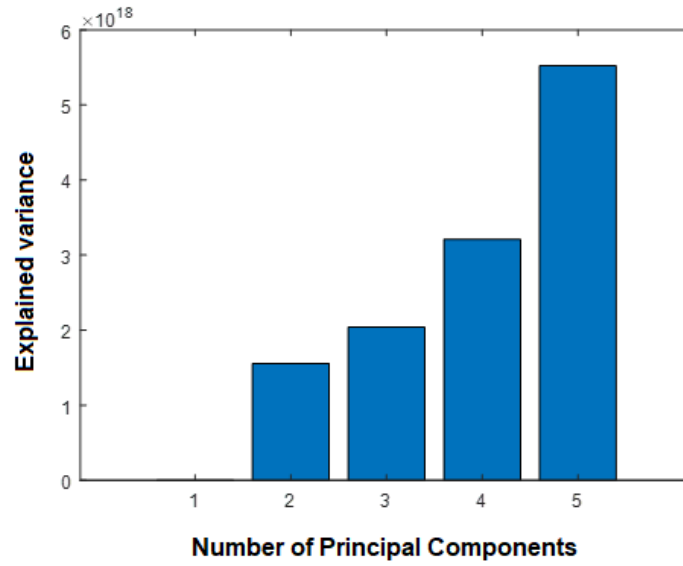
Figure 39 *The Principal Component Analysis (PCA) on the observation matrix X with 5 mixtures shows that only four of the five variables are significant. The individual explained variance of each principal component tells us how much information (variance) can be attributed to each PC and how many PC we are going to choose for the new feature subspace. Most of the variance can be explained by the first principal component alone. The second one still brings some information while the third and fourth ones explain less information about the distribution of the data. This suggests the number of the components to fit in the system, here four.*

Except for PALS, the other algorithms such as STALS, PALM and BC-VMFB require the choice of a thresholding parameter λ, for which five choices were tested, namely 0.01σ, 0.1σ, 1σ, 10σ and 100σ, where σ is always the standard deviation of the noise estimated from a signal-free region of *X*-matrix. These choices are made also for the wavelet-based versions of PALM and BC-VMFB. In terms of thresholding values, we can consider PALS as a special case of STALS with $\lambda_S = 0$.

On simulated data, we tested at the beginning all the algorithms with a fixed value of $\lambda_S = 10\sigma$. We obtained good results, especially with STALS and PALS (**Table 4**).

Table 4 Numerical indexes Amari, SIR and SDR (dB) (defined in I.2.2.7) of all the sources on simulated data of ^{1}H NMR spectra for all the algorithms with the thresholding parameter $\lambda_S=10\sigma$ (m stands for mean value).

	Algorithms					
	PALS	**STALS**	**PALM**	**BC-VMFB**	**PALM wav**	**BC-VMFBwav**
Amari	0.019	0.008	0.022	0.025	0.036	0.031
SIR(lim.)	26.7	52.4	22	24.4	20.7	22.2
SIR(ner.)	32.5	31.2	29.8	28.5	30.1	32.5
SIR(α-terp.)	19.5	45.7	41.6	25.2	24.3	23.5
SIR(β-car.)	47.6	29.3	19.9	24	21.4	21.7
SIR(m)	31.6	39.6	28.3	25.5	24.1	25
SDR(lim.)	26.7	51.4	21.7	24.4	20.4	21.3
SDR(ner.)	32.5	31.2	29.6	28.5	29.7	29.5
SDR(α-terp.)	19.5	44.9	40.8	25.2	24	22.8
SDR(β-car.)	47.4	28.6	19.6	23.9	21	20.6
SDR(m)	31.5	39	27.9	25.5	23.8	23.6

The estimated concentrations on simulated, with STALS at $\lambda_S =10\sigma$, are reported in **Table 5** where they are very close to the real ones. The numerical indexes SIR and SDR are ranged between 30 and 55 dB, then are considered very good.

Table 5 Estimated concentrations (mM) on simulated data of ^{1}H NMR spectra for STALS algorithm with the thresholding parameter $\lambda_S=10\sigma$.

	Limonene	**Nerol**	**α-Terpinolene**	**β-Caryophyllene**
Solution 1	23.33	25.82	9.12	10.88
Solution 2	17.05	12.02	15.51	14.99
Solution 3	9.17	14.12	18.53	4.69
Solution 4	20.71	7.00	13.49	11.96
Solution 5	5.06	9.06	10.88	13.13

Other simulations were performed using STALS with the five values of λ_S. To get a quantitative assessment, Amari, SIR and SDR are computed as shown in **Table 6** as well as the resulting spectra of limonene are reported in **Figure 40**.

Table 6 Numerical indexes Amari, SIR and SDR (dB) on simulated data of ^1H NMR spectra for STALS algorithm with five values of the thresholding parameter λ_S (m stands for the mean value).

	λ					
	0	**0.01σ**	**0.1σ**	**1σ**	**10σ**	**100σ**
Amari	0.019	0.019	0.019	0.019	0.008	0.018
SIR(lim.)	26.7	26.7	26.7	26.8	52.4	42.5
SIR(ner.)	32.5	32.3	32.2	31.9	31.2	27.6
SIR(α-terp.)	19.5	19.5	19.5	19.6	45.7	39.3
SIR(β-car.)	47.6	45.9	44.6	42.4	29.3	18.3
SDR(lim)	26.7	26.7	26.7	26.8	51.4	39.7
SDR(ner.)	32.5	32.3	32.2	31.9	31.2	27.3
SDR(α-terp.)	19.5	19.5	19.5	19.6	44.9	35.4
SDR(β-car.)	47.4	45.8	44.5	42.3	28.6	17.1
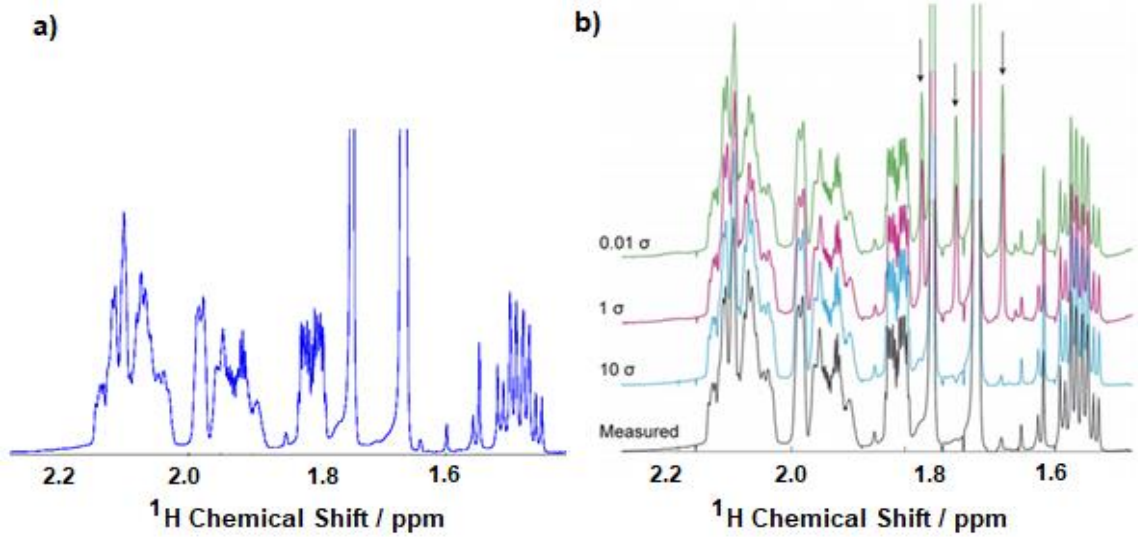
Figure 40 a) Limonene: (real) measured 1D ^1H spectrum of reference with a focus on the overlapped region. b) estimated ^1H NMR spectra with STALS, at different sigma and the measured spectrum of limonene. The arrows show the presence of three extra peaks. These peaks are not present for $\lambda_S=10\sigma$. These extra peaks are residual signals from nerol.

II.2.2. Real 1D Mixtures

Real mixtures are synthetic samples prepared by mixing the pure components, in which their concentrations vary within the samples. These analyzed mixtures *m* are five composed by four, *k*, sources of terpenes according to the rule $m \geq k$ mentioned before to estimate *k* from *m* (**Figure 41**).
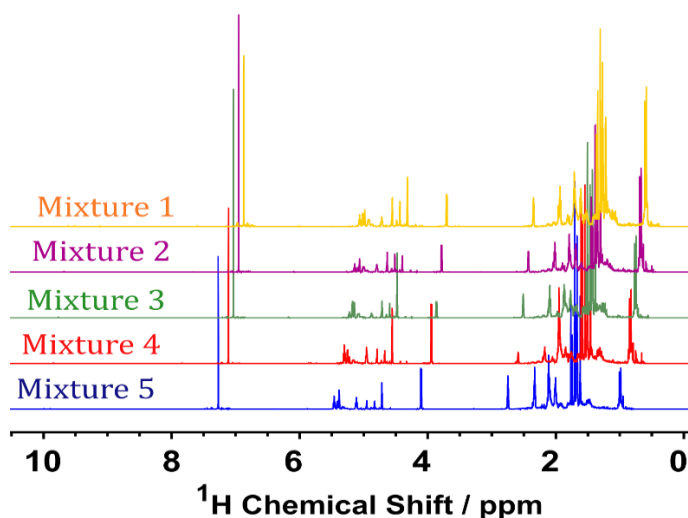
Figure 41 *The five real measured synthetic mixtures of the four terpenes composing the X-matrix, in which the concentration of each source varies from one sample to another. This means different intensities or integrals of the peaks belonging to the same molecule.*

The first attempt to decompose the real measured mixtures (*i.e.* the five rows of the *X*-matrix reported in **Figure 41**) failed. Despite the same calibration, the nature of 1D spectra can lead to variation in chemical shifts between spectra for same areas that have to be re-aligned. These shifts were corrected using the open source software *NMRProcFlow*, which provides a set of tools for processing and visualizing of 1D NMR spectra.[124] They can be due to changes in experimental conditions such as, instability of the instruments, temperature variations, that are harmful for the decomposition or statistical analysis.

II.2.3. Bucketing and Data Alignment on 1D NMR Spectra

Data alignment was necessary to explore the full information of the variance within a data set of NMR spectra of mixtures thus ensuring the estimation of the terpenes. It is a crucial step to assure the same resonance values from one sample to another along all the data set in consideration. The ^{1}H NMR spectra were directly exported in *NMRProcFlow* online software, and divided into 0.001 ppm-wide buckets leading to 10000 variables for 10 ppm total. Spectra

were divided in small regions which correspond to variables. A bucket table representing the *X*-matrix is then obtained and submitted to the alignment routine.

NMRProcFlow software aligns independently each NMR signal to a target or reference spectrum and it is based on a least-squares algorithm. Another alternative algorithm is that one developed by Savorani et *al.* called *Icoshift* (interval-correlation shifting) used on MATLAB (by MathWorks),[125] which maximizes the cross-correlation between user-defined intervals according to the peaks of the average spectrum. After the realignment, a comma-separated values .csv file, which allows data to be saved in a tabular format, is available to import in MATLAB. The solvent peak can be suppressed if necessary, using *NMRProcFlow*.

In **Figure 42**, the spectra of the real mixtures, which are the five rows of the *X*-matrix, and the simulated mixtures, the rows of X_{sim}, are reported before and after alignment.
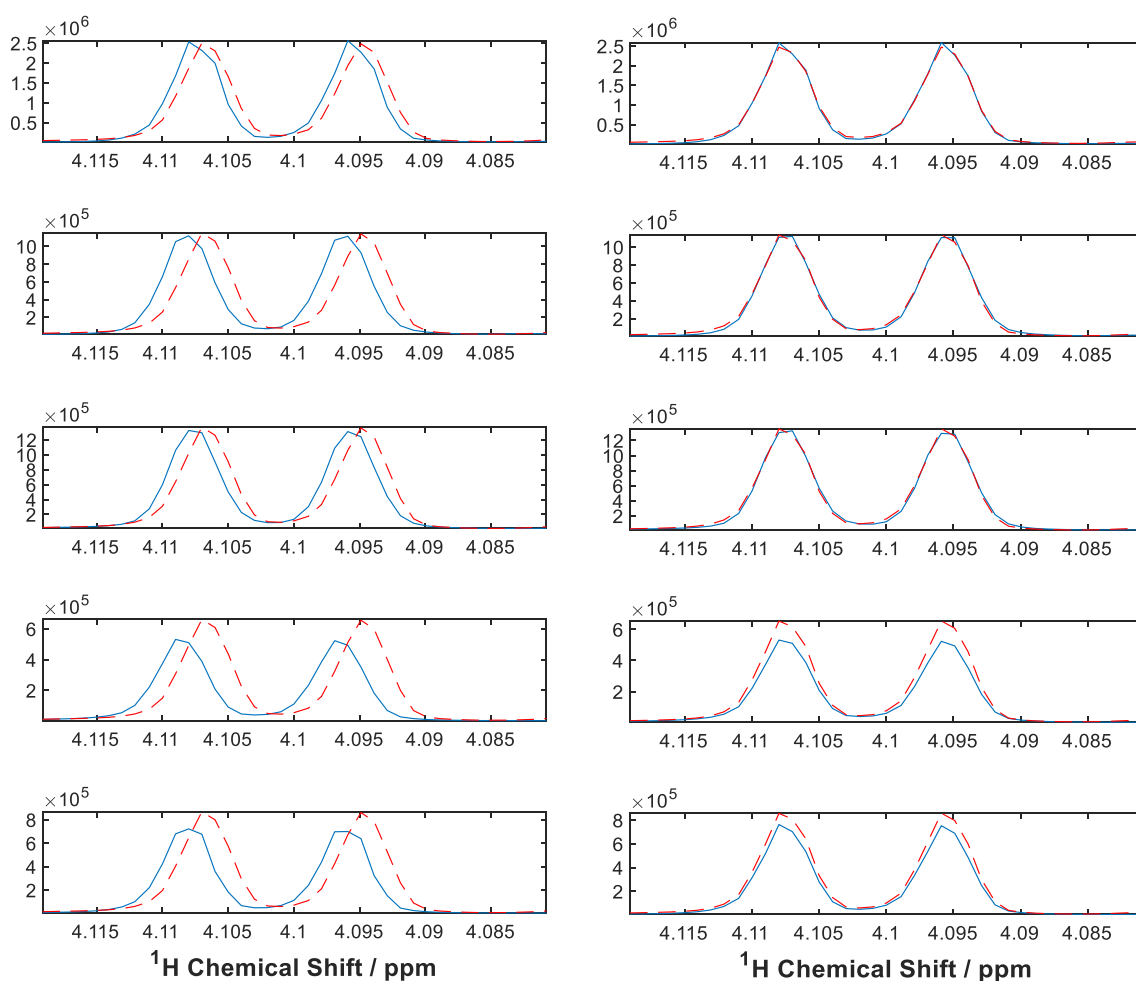
Figure 42 *Focus on a region of ^1H NMR spectra of X-matrix (in blue) and simulated X_{sim}-matrix (in red) before **a)**, and after alignment **b)**.*

As for the simulated data, numerical indexes are computed to assess the adequacy of the previous model and eventually to see differences between X and X_{sim}. SIR values of simulated data (**Table 7**), are fairly good at about 10 dB, indicating that the peaks are at the right position of chemical shift and that the number of spurious peaks is low, or even null. SDR values that compare real and simulated data sets are lower which may be interpreted as a consequence of signal amplitude differences. This suggests that a more adequate model to describe real mixtures should involve the correction of shifts and signals amplitude.

Table 7 Amari, mean SDR and mean SIR indices (dB) for comparing measured and simulated mixture spectra.

	Mixture 1	Mixture 2	Mixture 3	Mixture 4	Mixture 5
SDR_m	10.86	8.84	12.90	9.76	9.72
SIR_m	16.12	12.96	18.89	19.95	17.15

The de-mixing algorithms were then run on aligned real NMR mixtures, where the indexes using BC-VMFB with $\lambda_S = 1\sigma$ are reported in **Table 8**.

Table 8 Amari, mean SDR and mean SIR indices (dB) comparing the estimated and pure source spectra to evaluate the BC-VMFB algorithm on real aligned 1D mixture data, while the thresholding parameter λ=1σ.

	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Amari			0.081	
SDR_m	9.7	9.9	4.8	6.8
SIR_m	13.1	15	20.2	14.8

The estimated concentrations on aligned real 1D NMR mixtures are reported in **Table 9** with the relative errors in **Table 11** (estimated as in the example in **Table 10**).

Table 9 *Estimated concentrations (mM) on real aligned data of 1D ^1H NMR spectra for BC-VMFB algorithm with the thresholding parameter $\lambda_S=1\sigma$.*

	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Mixture 1	26.18	32.66	7.99	10.40
Mixture 2	17.79	11.13	15.67	18.94
Mixture 3	4.73	14.50	18.79	3.15
Mixture 4	24.42	2.07	13.75	10.63
Mixture 5	2.19	7.66	11.32	12.53

The relative errors in the estimated concentrations expressed in %, compared to ERETIC concentrations (**Table 2**), are obtained like in the following example (**Table 10**):

Table 10 *Example of calculation of relative errors in the estimated concentrations (in %) compared to the ERETIC concentrations (**Table 2**). Here the example for α-terpinolene in mixture 1.*

ORIGINAL CONC. (REFERENCE)	ESTIMATED CONCENTRATION
8.78 mM	7.99 mM
Relative Errors:	$\dfrac{(7.99 - 8.78)}{(8.78 : 100)} = \pm \%$

Table 11 *Relative errors in the estimated concentrations (in %) on real case of 1D ^1H NMR spectra using BC-VMFB with $\lambda_S=1\sigma$.*

	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Mixture 1	+12.34	+25.61	-9.03	-4.33
Mixture 2	+4.05	-6.68	+1.11	+26.27
Mixture 3	-47.70	+1.91	-0.53	-32.53
Mixture 4	+16.36	-69.78	+1.54	-11.15
Mixture 5	-55.03	-14.94	+4.74	-4.70

The following estimated sources on aligned real 1D data set are reported in **Figure 43**.
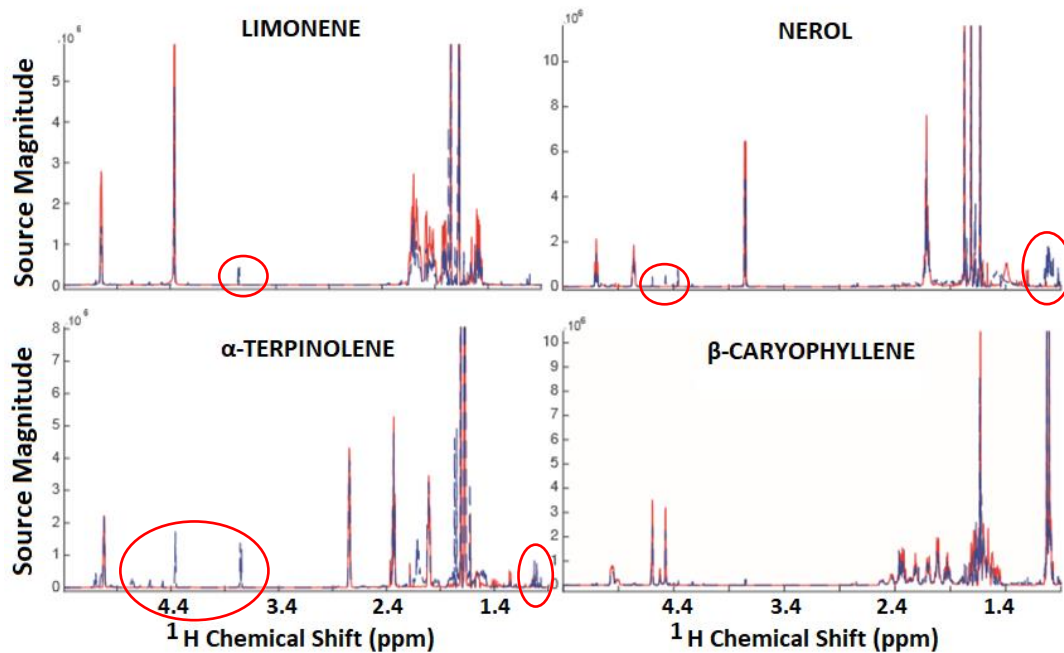
Figure 43 *^1H NMR spectra of the four terpene sources estimated using BC-VMFB algorithm with $\lambda_S=1\sigma$ (in blue), compared to the real spectra of reference (in red). Extra peaks are indicated by red circles.*

From these results, we can assess the relevance of the chosen mathematical indexes. The SIR (**Table 8**), which reflects the ratio between the estimated spectrum and the spurious peaks from the other sources (red circles) indicates that the α-terpinolene is the most "clean" spectrum whereas, visually, it is the one that contains the most "fake" peaks. At the same time, the limonene is indicated as the one containing the most spurious peaks while it is the least polluted. When we look at the peak intensities, we observe an important difference between the estimated and real sources for the limonene than for the α-terpinolene, that could explain the SIR values. however, the SDR appears to correlate well with the visual assessment and the presence of the spurious peaks.

II.2.4. Conclusions on 1D Data

The results on the 1D data show that pre-processing is a crucial step. In the case considered here, the alignment of the spectra turned out to be fundamental.[126] However, even after careful alignment, the Linear Instantaneous Mixture model can be incompletely satisfactory, showing significant departures from that model. Therefore, this may suggest that the model is not 100% adequate and that it would be worth considering more complex models, that could include, for instance, amplitude modulations as we observed on the terpene data we studied. One may also imagine including a spectral shift in the model, provided the phenomenon could be sufficiently well understood and therefore modelled. Despite these remarks, the major signals of each source spectrum are recovered, allowing the identification of every singular molecular fingerprint.

II.3. Results on 2D NMR Spectra of Terpenes

II.3.1. Results on 2D Simulated HSQC NMR Spectra

The algorithms were also evaluated on 2D ^{1}H-^{13}C HSQC on simulated and real data. The advantage in considering 2D NMR techniques for BSS is the spectral resolution, thus the possibilities of overlaps are lower and the requirement of source signal orthogonality (mutual uncorrelation) is easier to meet.[104,127] Moreover, the spectral dispersion in the ^{13}C dimension also provides the orthogonality constraint for the sources since their spectra to be separated do not overlap significantly. 2D HSQC spectra were recorded for the five mixtures. As discussed previously, since the acquisition conditions were similar, the difference in intensity of the correlation peaks is only due to the variation in the concentrations of the terpenes within the mixtures (**Figure 44**).
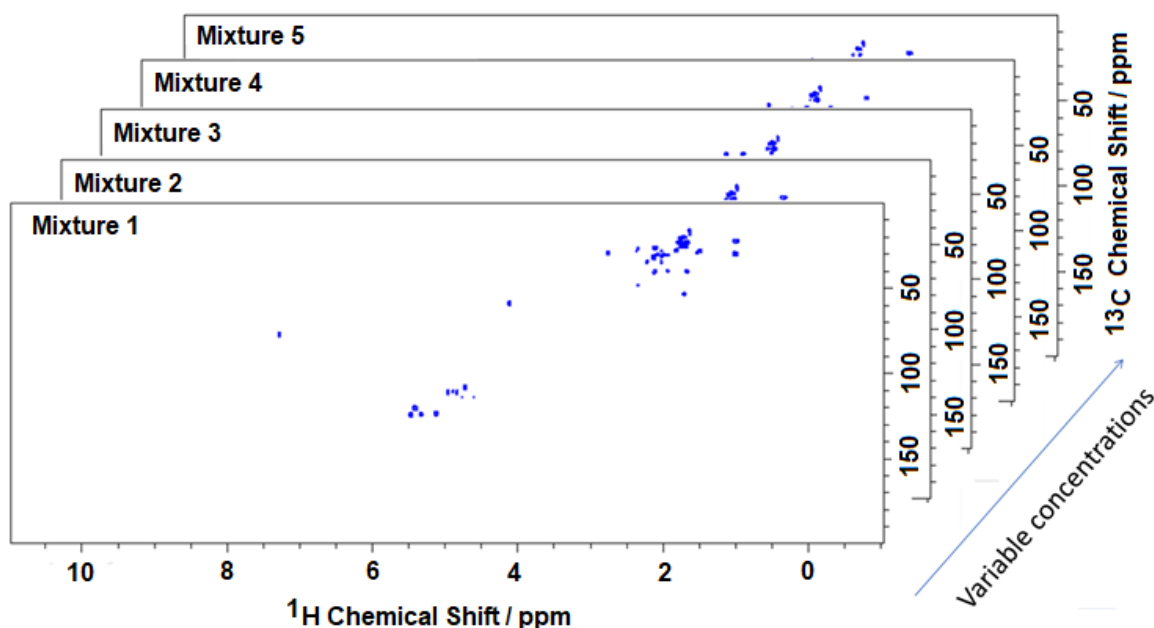
Figure 44 The five 2D HSQC spectra of the five synthetic terpene mixtures with different concentrations of each pure component, which means different intensities in correlation spots.

The first results concern simulated 2D data, using all the previous algorithms but with a higher thresholding parameter $\lambda_S=100\sigma$. We can assess the performances of the different algorithms using SIR and SDR (**Table 12**). The best results are obtained with STALS algorithm. The BC-VMFB algorithm, using wavelets, also leads to good results **Table 12** and **Figure 45**). The separation result is visually good and the fingerprint of each terpene is recovered.

Table 12 Numerical indexes SIR and SDR (dB) on simulated 2D ^1H-^{13}C HSQC NMR spectra for all the algorithms with the thresholding parameter $\lambda_S=100\sigma$ (m stands for the mean value).

	Algorithms					
	PALS	STALS	PALM	BCVMFB	PALM wav	BC-VMFB wav
Amari	0.135	0.012	0.015	0.11	0.081	0.016
SIR(lim.)	21.6	50.1	39.6	39.8	20.9	42.8
SIR(ner.)	12.4	39.1	31.7	10.3	21.8	34.6
SIR(α-terp.)	22	29	28	9.7	31.5	31.4
SIR(β-car.)	8.3	41.4	45.4	20.8	34.2	28.6
SIR(m)	16.1	39.9	36.2	20.2	27.1	34.4
SDR(lim.)	20.5	25.9	25.7	24.6	19.8	24.9
SDR(ner.)	11.6	16	17.2	9.3	16.4	16.4
SDR(α-terp.)	15.1	13.6	13.5	8.6	14.1	13.2
SDR(β-car.)	7.1	9.9	10.1	10.5	10.4	9.8
SDR(m)	13.6	16.3	16.6	13.3	15.2	16.1
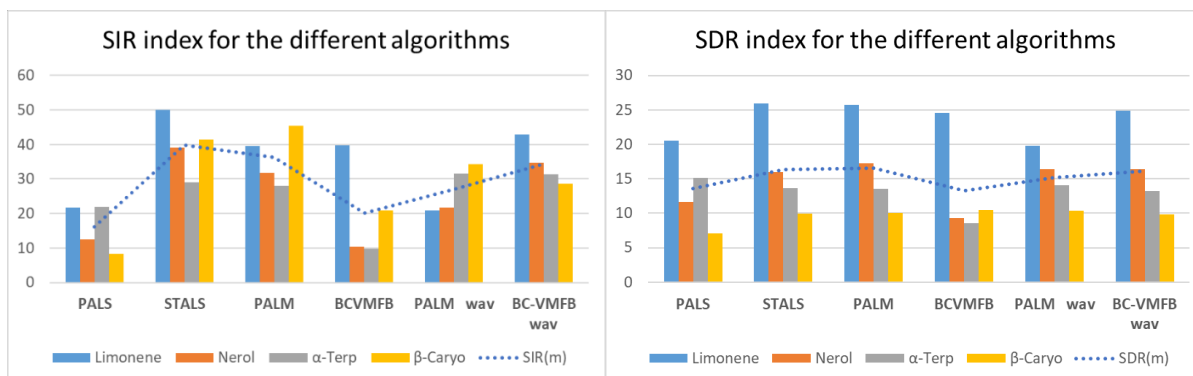
Figure 45 The estimated SIR (on the left) and SDR (on the right) indexes for all the sources using the different BSS algorithms.

II.3.2. Results on 2D Real HSQC NMR Spectra

The validated algorithms on 2D HSQC simulated data set were subsequently tested on real cases. The best performance was obtained as well with wavelet-based BC-VMFB algorithm using a $\lambda_S=10\sigma$, where the numerical evaluation indexes are reported in **Table 13**. SIR reported values are very good, but SDR estimated indexes are lower compared to the simulated 2D situation like in the 1D case. Again, this difference indicates that the linear mixing model, LIM, remains imperfect to well describe the real mixtures of components, even for the 2D spectra, when significant chemical shift changes are present due to intra- or intermolecular interactions.

Table 13 Amari index, mean SDR and mean SIR indices (dB) on real 2D HSQC data set for wavelet-based BC-VMFB with $\lambda_S=10\sigma$ (m stands for the mean value between all the sources).

	Limonene	**Nerol**	**α-Terpinolene**	**β-Caryophyllene**
Amari		0.042		
SDR$_m$	8.5	13.1	9.4	9.4
SIR$_m$	46.3	25.2	22.7	32.1

Despite the weaker numerical indexes compared to the simulated situation, the visual evaluation of the estimated sources leads to recognize the molecular fingerprints of these molecules as reported in **Figure 46, Figure 47, Figure 48, Figure 49** with the estimated HSQC spectra of limonene, nerol, α-terpinolene, β-caryophyllene, respectively. The crowded region between 1

and 2.5 ppm and the characteristic region at about 5 ppm are well recovered. The nerol has one spurious peak of β-caryophyllene at 1.1 ppm. The α-terpinolene has several extra peaks of β-caryophyllene especially at about 4.8 ppm, discussed in **Figure 50** below.
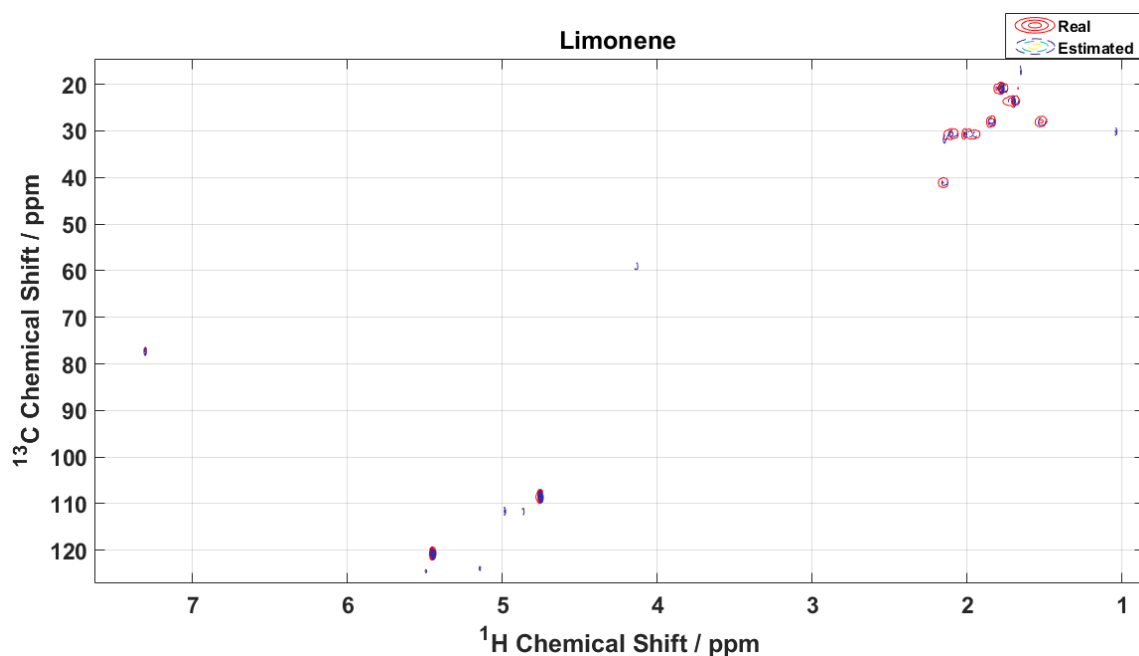
Figure 46 *Real 2D ^{1}H-^{13}C HSQC spectrum of the estimated limonene using wavelet-based BC-VMFB algorithm with λ=10σ (in blue), compared to the real spectra of reference (in red).*
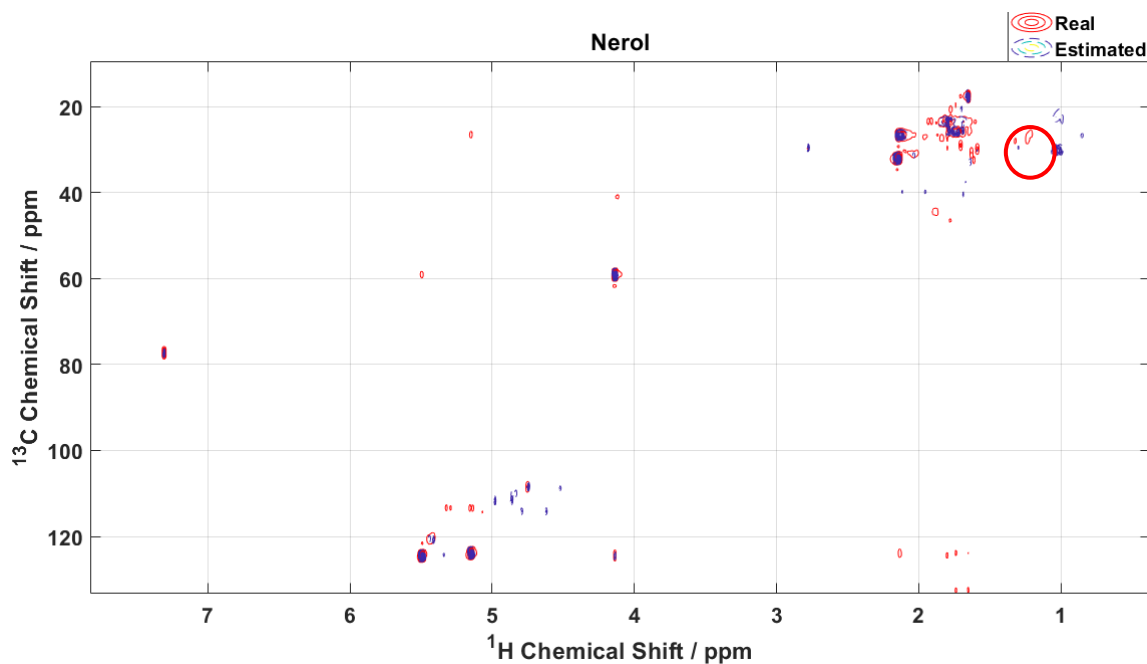
Figure 47 *Real 2D ^{1}H-^{13}C HSQC spectrum of the estimated nerol using wavelet-based BC-VMFB algorithm with $λ_s$=10σ (in blue), compared to the real spectra of reference (in red). Extra peaks are indicated by red circles.*
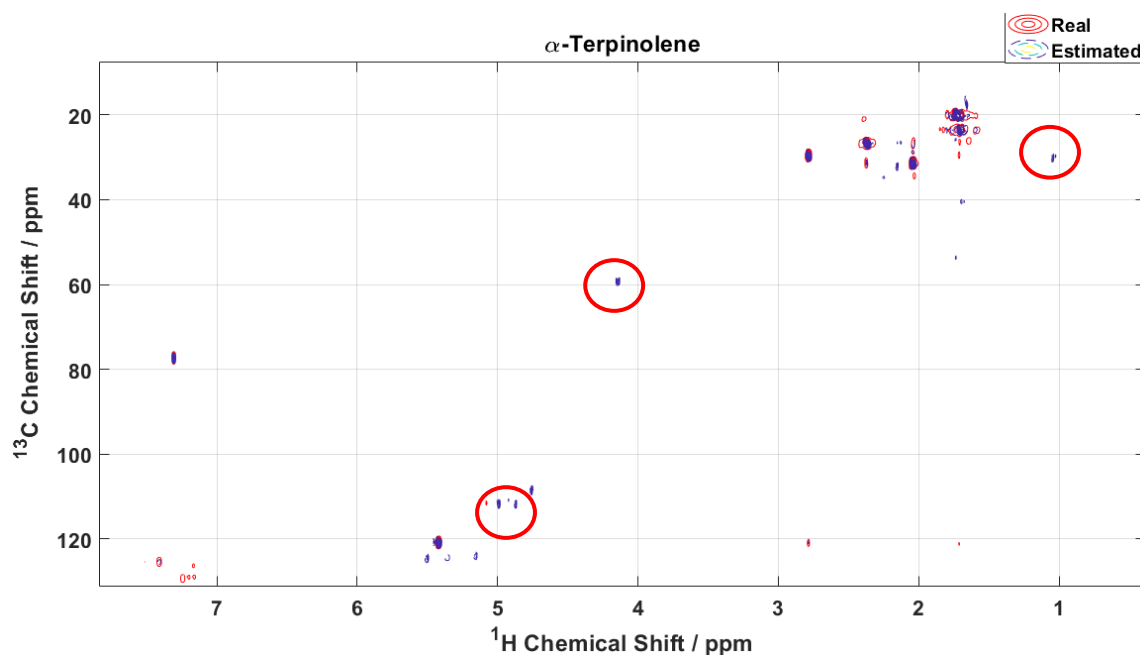
Figure 48 *Real 2D ^{1}H-^{13}C HSQC spectrum of the estimated α-terpinolene using wavelet-based BC-VMFB algorithm with λ=10σ (in blue), compared to the real spectra of reference (in red). Extra peaks are indicated by red circles.*
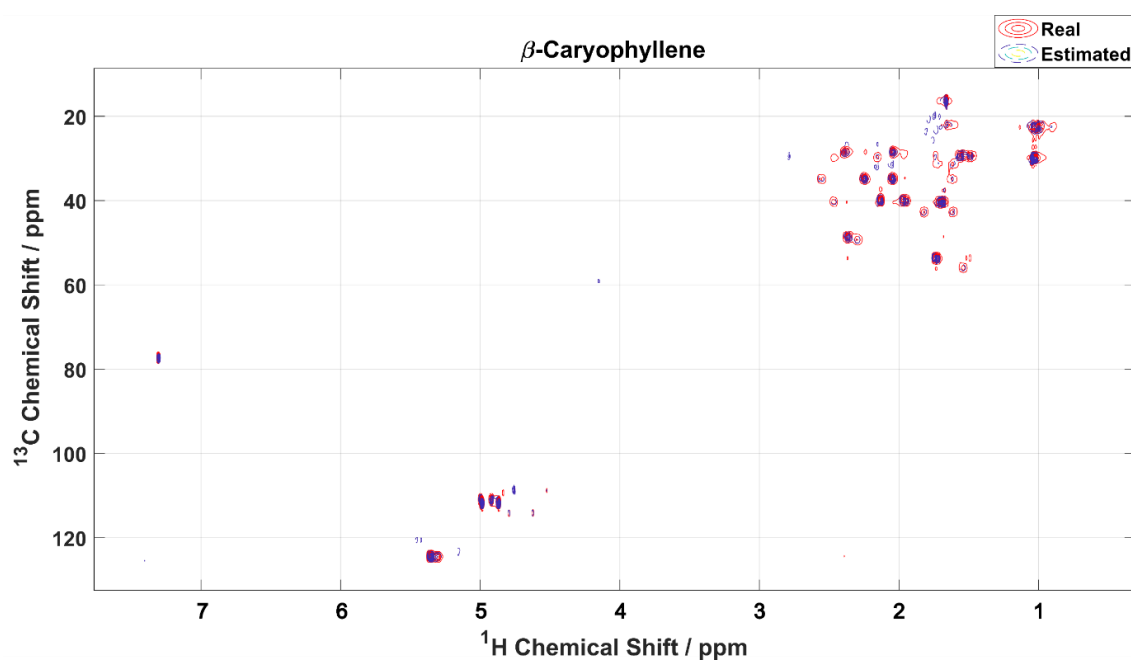
Figure 49 *Real 2D ^{1}H-^{13}C HSQC spectrum of the estimated source β-caryophyllene using wavelet-based BC-VMFB with λ=10σ (in blue), compared to the real spectrum of reference (in red).*

In **Figure 50**, a comparison between real 1D and 2D situations is shown. SIR is an indicator of the "pollution" from extra-peaks of other sources. These values are almost doubled passing to 2D resolved spectra, except for one source, the α-terpinolene where SIR remains constant. This

is visually detected on 1D and 2D spectra with some peaks belonging to the β-caryophyllene at 4.8 ppm and 1 ppm, and to the nerol at 4.1 ppm.
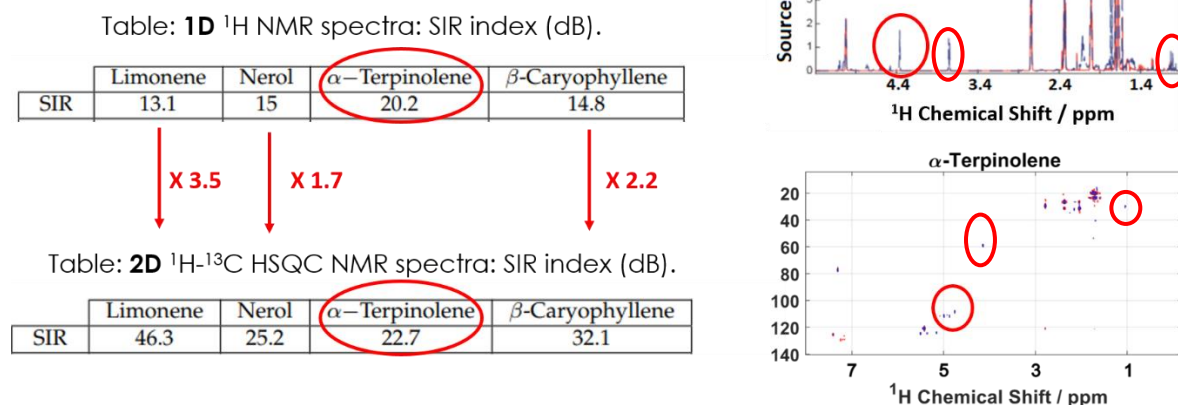
Figure 50 *Comparison of SIR indexes in 1D and 2D real data sets with the corresponding spectra of the α-terpinolene.*

The estimated concentrations on the real 2D dataset (**Table 14**), with the relative errors compared to ERETIC concentrations (**Table 2)** are reported in **Table 15**. The quantitative estimation here is better than the one-dimensional real situation (**Table 8** and **Table 9).**

Table 14 *Estimated concentrations (mM) on real 2D NMR spectra for wavelet-based BC-VMFB algorithm with the thresholding parameter $λ_s=10σ$.*

	Limonene	**Nerol**	**α-Terpinolene**	**β-Caryophyllene**
Mixture 1	23.25	30.03	7.38	10.90
Mixture 2	17.26	9.95	14.54	14.97
Mixture 3	8.62	16.78	23.48	4.74
Mixture 4	21.83	4.93	14.08	12.09
Mixture 5	4.36	6.35	8.04	12.95

Table 15 *Relative errors in the estimated concentrations (in %) on real case of 2D NMR spectra using wavelet-based BC-VMFB with $λ_s=10σ$. Amari index is equal to 0.042.*

	Limonene	**Nerol**	**α-Terpinolene**	**β-Caryophyllene**
Mixture 1	-0.39	3.16	-7.02	9.53
Mixture 2	-8.56	-3.43	-2.66	10.90
Mixture 3	5.84	8.42	2.94	6.68
Mixture 4	5.74	-14.40	3.47	-12.84
Mixture 5	-10.66	-29.52	-25.62	-1.52

The non-quantitative response due to the dependence of the correlation peak volumes on the efficiency of the magnetization transfer of the INEPT delays in the classical HSQC, depending on the heteronuclear coupling constant (J_{IS}), is not important in this study for the application of the BSS. As said before, the changes in the peak intensities within the mixtures are only due to variations in concentrations detected by the BSS.

II.3.4. Discussions on 2D HSQC results

For two-dimensional data set of terpenes,[126] we observed that alignment, as a pre-processing step, was not problematic like for one-dimensional spectra. However, it must be noticed that it was necessary to re-scale the intensities using "nc_proc" processing parameter to normalise all the 2D spectra.

The recovered molecular fingerprints after the BSS and therefore the identification, as well as the estimated concentrations, are of better quality in 2D than the 1D case as shown visually in the histograms in **Figure 51**. Some exceptions are present, in fact, in some cases, the estimated concentrations (expressed in mM) are better in 1D spectra such as for nerol in mixture 1, β-caryophyllene in mixture 3 and mixture 2, nerol in mixture 4, and limonene in mixture 5.
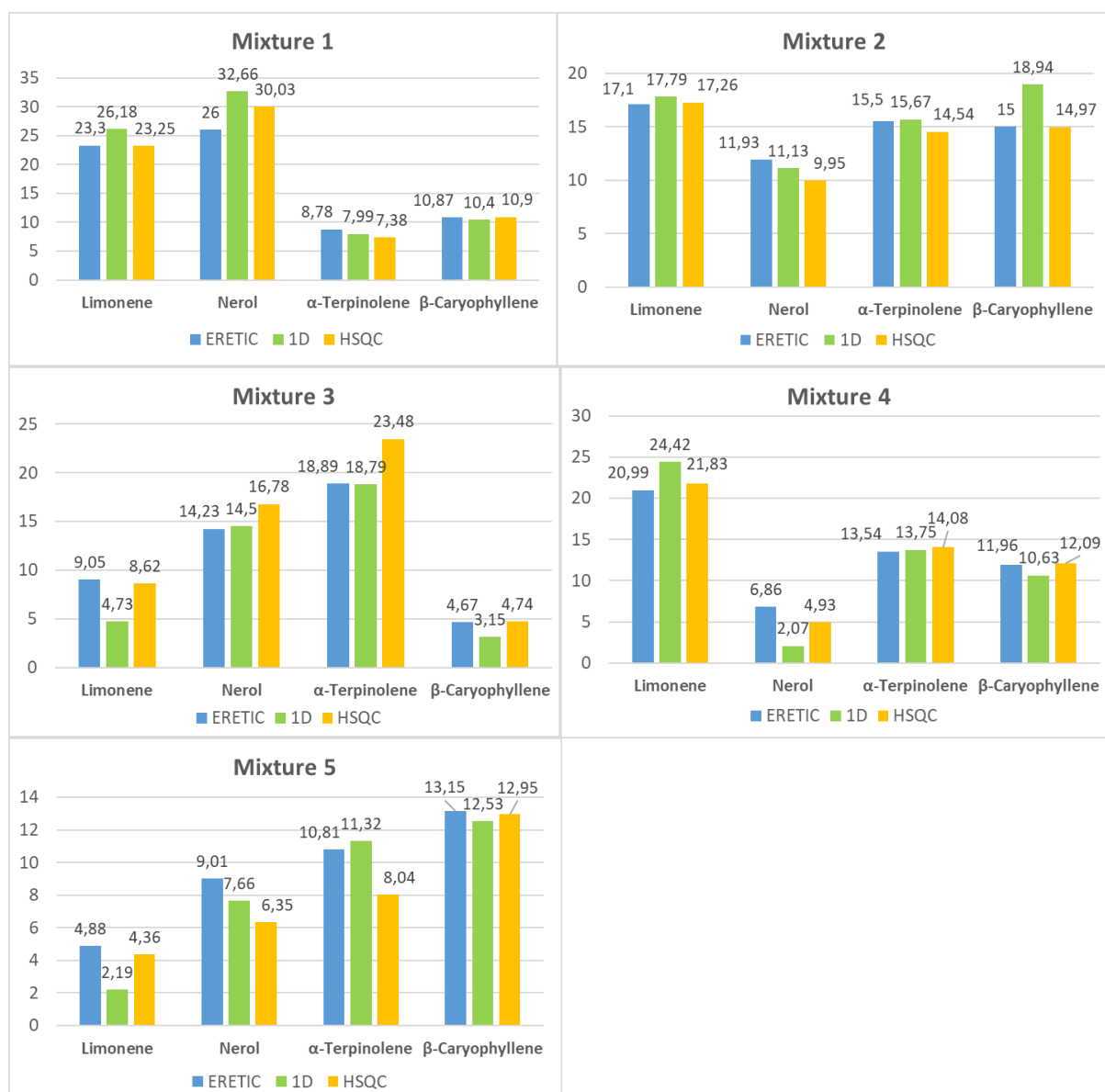
Figure 51 *Histograms to compare visually the ERETIC2 real concentrations (mM) of reference to the estimated 1D and 2D HSQC concentrations (mM) on real mixtures after the BSS.*

Histograms representing the relative errors (compared to the ERETIC2 concentrations) in the estimated concentrations, expressed in %, are reported in **Figure 52** for the real cases of 1D proton and 2D HSQC NMR. The relative errors in the 2D case are smaller than the 1D situation as confirmed with the previously estimated concentrations which are better estimated (**Figure 52**).

Figure 52 *Histograms with the relative errors in the estimated concentrations (in %) in 1D and 2D data set for terpene mixtures compared to the ERETIC2 concentrations of* **Table 2**. *In the 2D case these errors are smaller than the 1D, consequence of the better estimated concentrations (mM).*

This better estimation may be interpreted as a consequence of data sparsity adding spectral dimensions. Everything is confirmed by the relatively high values of SIR and SDR indexes, and also by the Amari values. As seen, SDR values are lower which indicates the presence of distortions. However, in our work, we always used rather the SIR measurements to estimate the quality of estimated source spectra.

The computational workload is significantly high using two-dimensional spectra, passing from less than 10 minutes in 1D to about three hours in 2D case. Moreover, it increases using higher-dimensional spectra (tested on decomposing 3D spectra during this thesis, not shown). NMR

spectra, to be red and created in the code, are represented by matrices that will become bigger using more dimensions. This constitutes a limit, but it is obvious that the performances in 2D are better.

Non-convex minimization problems leading to local minima require to optimize the objective function with an initialization. The algorithms are very sensitive to initialization where the BSS requires the knowledge of the number of sources to fit (S_0) and a mixing matrix of departure (A_0) to start the computation.

These results indicate that the BSS algorithms are well suited for samples whose compositions and concentrations are relatively known, as is the case for the most studied samples, in metabolomics for example, where the composition is known and the concentrations are ranged in well-specific intervals. However, for the unknown samples, more developments are needed to perform the completely blind source separation.

II.3.5. Results on 2D HSQC NMR Spectra Using Non Uniform Sampling

The application of BSS algorithms to decompose Non Uniform Sampled 2D HSQC spectra of the five mixtures of the four terpenes is evaluated in this section. The impact of three different compression factors of 4, 8, and 16, respectively using only 25%, 12.5%, and 6.25% of the acquired t_1 increments, was assessed on the performances of the algorithms, in terms of separation capability and quantitation.

The first tests were carried out on simulated mixtures, with the more difficult case with NUS=6.25%. The best estimation of the A matrix and the sources was achieved using PALM algorithm with $\lambda_S = 10\sigma$. The SDR of β-caryophyllene, was the lowest value among the other sources, which could be explained by the higher level of noise present in this spectrum. we chose therefore to apply a $\lambda_s = 10\sigma$ for PALM calculations of the other NUS data and the

obtained results are reported in **Table 16**. The Amari index, SDR, and SIR were better for NUS 25% and the quality of the de-mixing process decreased until NUS 6.25%.

Table 16 *^1H–^{13}C NUS HSQC spectra: numerical results on simulated data for PALM algorithm with λs=10 σ at NUS 25% (**a**), 12.5% (**b**) and 6.25% (**c**).*

a)

HSQC NUS=25% λ=10σ	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Amari	0.019			
SDR	27.27	20.25	18.28	14.49
SIR	36.88	29.35	33.25	41.72

b)

HSQC NUS=12.5% λ=10σ	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Amari	0.0204			
SDR	26.07	20.89	19.16	15.18
SIR	36.29	32.08	34.89	27.39

c)

HSQC NUS=6.25% λ=10σ	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Amari	0.0234			
SDR	24.89	18.78	17.72	16.02
SIR	35.59	29.24	43.04	28.18

The second tests were carried out on real mixtures spectra always using the five values of $λ_s$, and σ was calculated for each data set. The preliminary tests were done on the harder situation with a compression factor of 16 (6.25% of indirect sampled points recorded).

In **Table 17** the numerical results on real data were reported, where the best performances are obtained with PALM algorithm with $λ_s=10$ σ. Even if the quality indices are worse than the simulated case, they are still good. With BC-VMFB the best result is obtained with $λ_s=100$ σ. Between the two performances, the Amari index is better with PALM, confirmed also by the better-estimated mixing matrix *A*, here not reported.

Table 17 ^1H-^{13}C *HSQC NMR Spectra: numerical results on real mixtures, using λ_s =10 σ for PALM and λs=100 σ for BC-VMFB (m stands for the mean value).*

Real HSQC, NUS 6.25%		
	PALM λ=10σ	BC-VMFB λ=100σ
Amari	**0.1066**	**0.2460**
SIR(lim.)	30.9	27.5
SIR(ner.)	16.5	23
SIR(α-terp.)	10.8	15.1
SIR(β-car.)	8	8.1
SIR(m)	**16.5**	**18.4**
SDR(lim.)	4.4	2.8
SDR(ner.)	7.9	8.7
SDR(α-terp.)	6.9	8.6
SDR(β-car.)	4.2	3.4
SDR (m)	**5.9**	**6**

In **Table 18**, the results of the PALM algorithm are reported using different values of the thresholding parameter.

Table 18 ^1H-^{13}C *HSQC NMR spectra at NUS = 6.25%: numerical results on real mixtures, using different λ_s for PALM algorithm.*

PALM with λ_S →	0.01 σ	0.1 σ	1 σ	10 σ
Amari	**0.2480**	**0.2271**	**0.2608**	**0.1066**
SIR(lim.)	6.2	7.2	7.9	30.9
SIR(ner.)	2.4	3.4	2.7	16.5
SIR(α-terp.)	3.2	3.9	3.7	10.8
SIR(β-car.)	2.5	3.2	3.1	8
SIR(m)	**3.6**	**4.4**	**4.3**	**16.5**
SDR(lim.)	2.7	3.2	3.5	4.4
SDR(ner.)	0.5	1.3	0.7	7.9
SDR(α-terp.)	1.7	2.3	2.1	6.9
SDR(β-car.)	0.7	1.2	1.1	4.2
SDR (m)	**1.4**	**2**	**1.8**	**5.9**

The best result is obtained with PALM algorithm using λ_s = 10 σ as reported in **Table 17** and **Table 18**. SIR values are acceptable and indicate that peaks are located at the right place in their ppm value and have not many interferences of other pure components. Contrarily, the SDR values are lower, generally related to the noise, and usually interpreted as a consequence of amplitude modulations. Here with NUS, we can suppose that these lower indexes are a

consequence of the t_1 noise. These spectra with NUS=6.25% and λ_s =10 σ, are plotted, in red, with the real data of reference and, in blue, the estimated data after the BSS (**Figure 53**). The zooms of different regions show that the molecular fingerprints are recovered for each source.
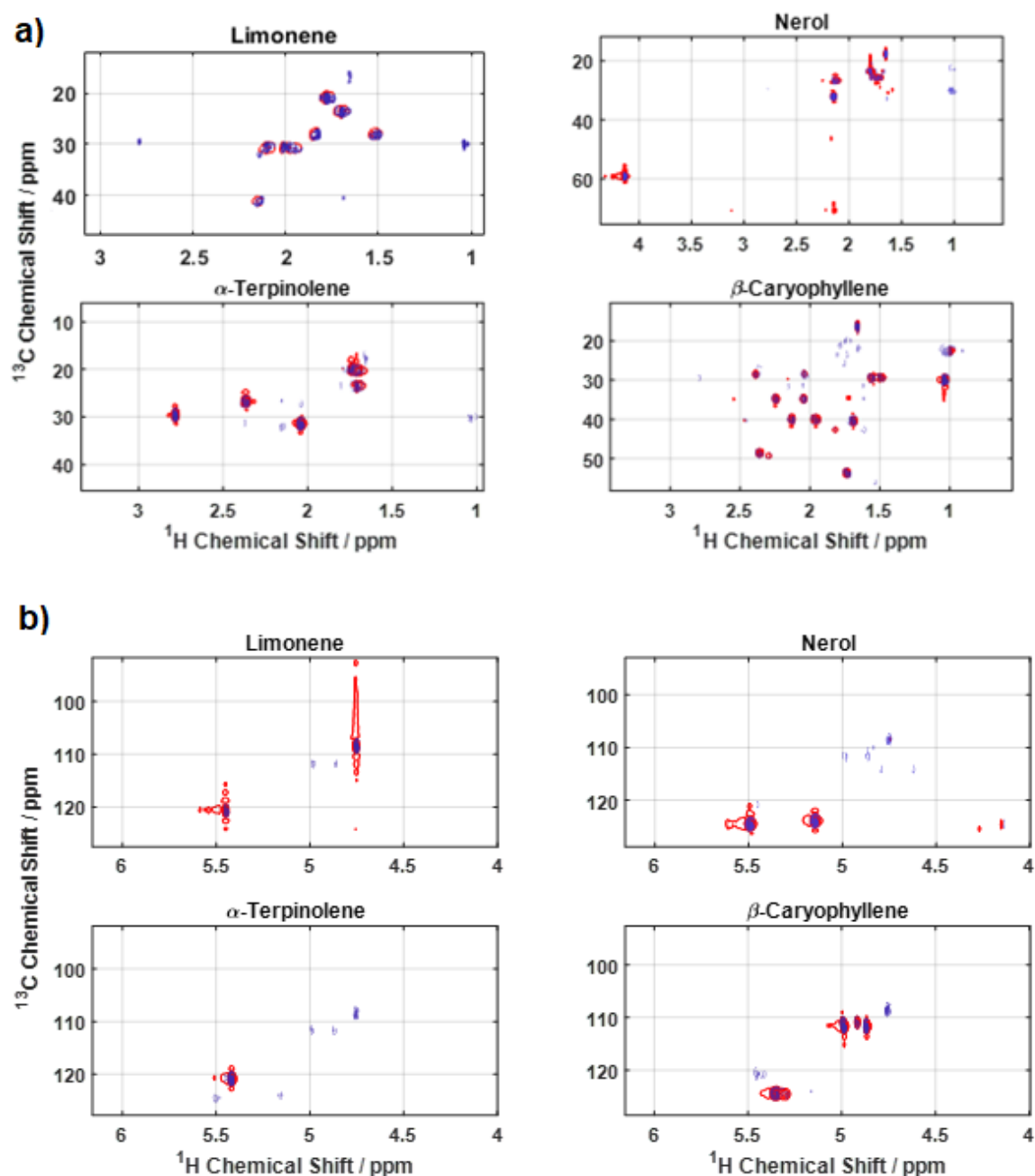
Figure 53 *- ^{1}H-^{13}C HSQC recovered spectra with NUS at 6.25% after BSS with PALM algorithm, for λ_s = 10 σ. In blue, the estimated source spectra and, in red, the real data. a) On the top, there are four zooms of the common overlapped region of the terpenes (1-2.5 ppm). **b)** On the bottom, there are four zooms in the isolated region characteristic for each terpene (4.5-5.5 ppm).*

Interesting is the comparison using different contour levels to plot the estimated sources. **Figure 54** depicts the superposition of estimated (blue) and real NUS-6.25% (red) HSQC spectrum of β-caryophyllene, with two different levels (lower level to the right, and higher-level to the left). With the low contour level, we highlight the presence of downsampling artefacts that seem to be filtered out in the estimated spectra (blue). We can attempt an explanation by considering that the downsampling artefacts do not behave like real NMR signals from one mixture to another, so the BSS cannot link them to the source. We discover here an unexpected virtue of the BSS which can restore artefact-free pure spectra from data set of mixtures 2D sub-sampled spectra with very high compression factors.
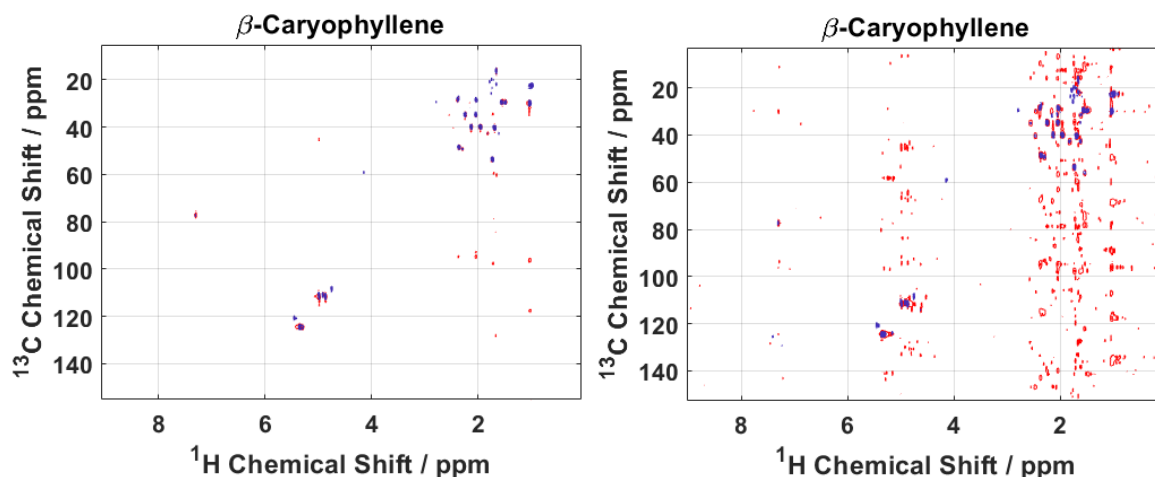
Figure 54 ^1H-^{13}C HSQC NMR recovered spectra of β-caryophyllene with NUS at 6.25% after BSS with PALM algorithm, for λ_s=10 σ. In blue, the estimated source spectra and, in red, the real data. On the left with a higher contour level and on the right with a lower one used to plot them.

In the end, tests with the NUS using 25% of the indirect sampled points are performed, confirming the expectation in which more peaks, almost all, are recovered than with NUS at 6.25%. The SIR and SDR values are better here of almost five units more than the ones at 6.25%. Performances using PALM with λ_S=10σ are the best (**Table 19**) and in general, we privilege the situation which estimates the higher average SIR index.

Table 19 ^1H-^{13}C HSQC NMR spectra at NUS = 25%: numerical results on real mixtures, using $λ_s$ =10σ for PALM algorithm (m stands for the mean value).

PALM	$λ_s$=10 σ
Amari	**0.0567**
SIR(lim.)	36.7
SIR(ner.)	18.1
SIR(α-terp.)	17.7
SIR(β-car.)	11.7
SIR(m)	**21.1**
SDR(lim.)	5.9
SDR(ner.)	13
SDR(α-terp.)	10.2
SDR(β-car.)	8.3
SDR (m)	**9.3**

Regarding the quantitative aspect, using NUS=6.25% the estimated concentrations are very similar but only a little bit worse than the ones at NUS=25% reported in **Table 20** with their relative errors in

Table 21.

*Table 20 Estimated concentrations using PALM for real case of HSQC spectra with $λ_s$ = 10σ with NUS at 25%. The symbol * indicates misestimated concentrations (mM).*

$λ_S$ = 10 σ NUS 25%	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Mixture 1	25.8	30.2	6.9	11
Mixture 2	14.4	10.2	15.9	17.7*
Mixture 3	8	14.4	19.7	4.1
Mixture 4	24.2	4.8*	14.4	10.9
Mixture 5	3*	8.3	10.6	12.2

*Table 21 Relative errors in the estimated concentrations (in %) on real case of HSQC spectra with NUS at 25% using PALM with λ=10σ. The symbol * indicates misestimated concentrations.*

$λ_S$ = 10 σ NUS 6.25%	Limonene	Nerol	α-Terpinolene	β-Caryophyllene
Mixture 1	+10.73	+16.15	-21.41	+1.2
Mixture 2	-15.79	-14.5	+2.58	+18*
Mixture 3	-13.13	+1.19	+4.28	-12.2
Mixture 4	+15.29	-30*	+6.35	-8.86
Mixture 5	-38.52*	-7.88	-1.94	-7.22

To sum up, these results demonstrate that the BSS works well to recover NUS spectra in simulated and real mixtures, also if, on the real data sets, the evaluation indexes are much smaller than the simulated case. The t_1 noise is not present in the spectra but some peaks, also with high intensities, are not recovered. However, always in the real case, the performances are quite good for limonene and β-caryophyllene, but for the nerol and α-terpinolene, there are some differences from the real data, especially with the compressing factor of 16 (**Figure 53**). The plotted spectra show that the molecular fingerprints are recovered sufficiently.

We observed that the BSS recovered spectra are similar to the original ones of the pure terpenes with less t_1 noise in general.

It must be noticed that by using multidimensional NMR spectra, the computational workload increases. Increasing the number of spectral dimensions, the matrices representing the NMR spectra are bigger. Since the reconstruction of the non-sampled points is performed prior to Fourier transform, the dimension of *X*-matrix is like that one in the US spectra. Thus, the NUS spectra have no benefit on the computational workload.

II.3.6. Results on 2D TOCSY Spectra

This short paragraph is dedicated to the assessment of the application of BSS on TOCSY data sets. This sequence is often used in the analysis of complex mixtures such as metabolomics for example.[128–132]

Un-mixing algorithms were run on real mixture spectra using five values of λ_s, where σ was calculated for the terpene TOCSY data set. The best results were obtained using BC-VMFB with $\lambda_s=10\sigma$.

In **Table 22** the estimated sources after BSS in the column on the left are compared to the acquired sources on the right.

Table 22 Estimated TOCSY sources after BC-VMFB algorithm using $\lambda_s=10\sigma$ on the left and the real acquired sources on the right.

ESTIMATED SOURCES	REAL SOURCES OF REFEENCE

As it is shown in **Table 22**, almost all the molecular fingerprints are visually recovered fairly good, also if some spurious peaks are present sometimes especially for the estimated limonene, and α-terpinolene in the region between 4 and 5 ppm. The numerical indexes reported in **Table 23** show the mean SIR value, related to the number of extra peaks of other sources, and the SDR value are not so high. Amari index related to the estimated mixing matrix *A* of the concentrations denoting that the estimated concentrations are quite different from the real matrix *A*. The estimated concentrations are in fact not satisfying and here are not reported.

Table 23 *2D TOCSY NMR spectra with NUS = 25%: numerical results on real mixtures, using λ_s =10σ for BC-VMFB algorithm where m stands for the mean value.*

Real 2D TOCSY NMR spectra with NUS = 25%	
Amari	0.406
SIR m	7.6
SDR m	2.4

The low SIR and SDR values could be improved working on the initialization step, here carried out with the JADE algorithm.

The future development of TOCSY spectra can be done by removing the diagonal peaks which are more intense than the cross-peaks, which can constitute an additional difficulty for the decomposition of the spectra. Another option could consist in not suppressing the diagonal because it is generally very precious for the structural analysis, therefore pre-processing could be proposed to reduce the diagonal rather than suppress it.

II.4. BSS on Data Sets with Spectroscopically Induced Variations (Diffusion Spectroscopy)

As it has been shown, the BSS algorithms decompose the spectra of mixtures using the concomitant variations of the peaks belonging to the same molecule within the different mixtures. In this part, we consider the use of "artificial" intensity variations, induced by spectroscopic methods that mime the concentration variations. With such approach, BSS algorithms could be used even if only one mixture sample is available. This situation was dealt previously by J. M. Nuzillard and co-workers[133] to demonstrate the LP-BSS (Linear Programming BSS) method on a mixture of menthol and β-sitosterol submitted to PFGSE experiment. In this case, the diffusion-modulated spectra can be decomposed in the same way to mixtures which are modulated by the differences in concentrations.

To evaluate the algorithms studied in this thesis in such a situation, we used the data sets of the NMR diffusion experiment of different samples to decompose NMR spectra blindly, without prior knowledge of the signal evolution, unlike other diffusion processing methods. In this work, all algorithms presented in the previous sections were tested, with X containing the set of spectra evolving as a function of the gradients. The best results were obtained using wavelet-based BC-VMFB mentioned above, with the thresholding parameter $\lambda_S = 0,01\sigma$. The results are assessed qualitatively and the mathematical indexes are not presented.

Standard DOSY spectra ("dstegp3s" according to Bruker nomenclature) were recorded by a conventional convection compensated pulse sequence, based on the stimulated echo and incorporated bipolar gradient pulses, and an Eddy current delay (BPP-LED). The shape of all gradient pulses was Smoothed Square and the LED delay was 5 ms. The gradient strength, g, was linearly incremented in 16 steps from 2% to 98% of its maximum value, and 16 scans were recorded. The diffusion delay and gradient pulse duration were specifically adjusted for each studied system in this section.

II.4.1. 1D and 2D DOSY experiments

II.4.1.1. Results on terpene sample

The first test was applied using a home-made 3D DOSY-HSQC pulse sequence obtained by concatenating the "dstbpgp3s" and "hsqcetgpsi3d" sequences, the time domains have been assigned as follows: td1 diffusion, td2 indirect HSQC dimension, td3 direct dimension. Since this experiment gives a 3D spectrum, there are three axes that give three information for each one: the chemical shift of the proton, that one of the carbon, and the diffusion dimension in which the gradient strength of the magnetic field varies. The result can also be seen as various 2D HSQC spectra acquired for each value of gradient strength along the diffusion dimension, in which the intensities of the cross-peaks decrease. The BSS was tested using all the 2D HSQC spectra, considered as mixtures. The resolution due to translational mobility has to be quite sufficient to allow the detection of the four pure compounds. Preliminary data showed that BSS algorithms do not work on terpene mixtures. It showed that there is a sensitivity to the diffusion coefficient resolution. The difference in the molecular diffusion coefficients has to be large enough.

To improve this resolution, we resorted to a so-called "chromatographic-NMR" approach which consists in using a solid matrix to better differentiate the diffusion coefficients thanks to the difference in affinity of the molecules to the solid matrix and the different interaction of the compounds with the solid and the liquid phase which can affect the molecular movements.[58,134,135]

The experiments were carried out with a HR-MAS (High Resolution-Magic Angle Spinning) probe-head spectrometer. 50 µl of the four terpenes solution dissolved in $CDCl_3$ was introduced into a rotor (4 mm) filled to 2/3 with silica, and before closing the rotor, silica was added to equilibrate the sample. The used LiChrospher Si 100, purchased from Merck, is a sorbent that

offers optimal polar properties for normal-phase separations. It is based on spherical silica

particles with a pore diameter of 100 Å, and a particle size of 10 µm.

In **Table 24**, diffusion coefficients after chromatographic NMR are reported which are more

differentiated from that one obtained with liquid state analysis.

Table 24 Measured diffusion coefficients of terpenes with a DOSY experiment using a 600 MHz liquid state spectrometer compared to the ones obtained after chromatographic-NMR with a normal solid phase and the apolar solvent CDCl$_3$ using a 400 MHz HR-MAS spectrometer at the spinning rate v_R = 4000 Hz.

Diffusion coefficient m^2/s	Liquid state	After Chr-NMR (in CDCl$_3$)
LIMONENE	$1,759 \times 10^{-9}$	$5,05 \times 10^{-9}$
α-TERPINOLENE	$1,592 \times 10^{-9}$	4×10^{-9}
β-CARYOPHYLLENE	$1,577 \times 10^{-9}$	$3,58 \times 10^{-9}$
NEROL	$1,616 \times 10^{-9}$	$1,50 \times 10^{-9}$

Results on BSS applied on 1D DOSY of one mixture of terpenes after Chr-NMR do not allow

to recover the sources spectra, probably due to the non-sufficient diffusion resolution and an

important number of the sources to be estimated.

II.4.1.2. Results on polysaccharide sample

The successive experiments were performed on a synthetic mixture of three sugars,

particularly, a mono-saccharide, mannitol, a disaccharide, sucrose, and a pentasaccharide,

maltopentaose (**Figure 55**).
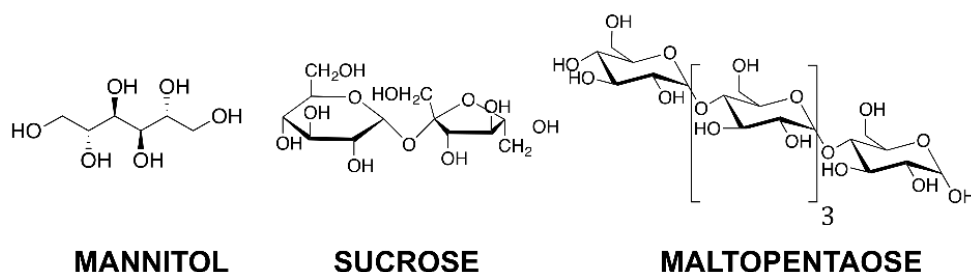
Figure 55 The three saccharides constituting the sample of polysaccharide mixture: mannitol (mon-saccharide), sucrose (disaccharide), maltopentaose (pentasaccharide).

The idea was to have more different molecular sizes and therefore more differentiated diffusion coefficients. As shown in **Figure 56**, they have a superposed region on the spectrum between 3 and 4 ppm, which represents another challenge, and a characteristic region between 4 and 5.5 ppm.
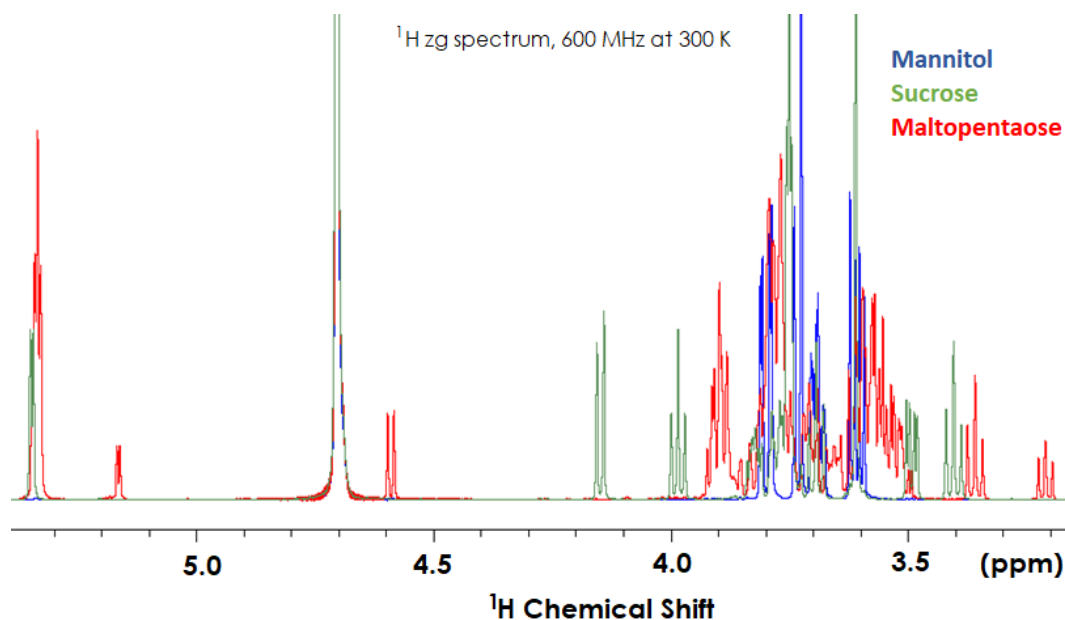
Figure 56 The NMR spectrum of mannitol in blue, in green the sucrose and in red the maltopentaose with the overlapped region at about 3.7 ppm.

Using the 1D spectra from a classical 2D DOSY experience, the separation results were not good enough because the resolution in diffusion was not sufficient enough (**Table 25** and **Figure 64**).

Table 25 Measured diffusion coefficients from DOSY experiment of the pure components in the mixture of polysaccharides.

POLYSACCHARIDE MIXTURE	DIFFUSION COEFFICIENTS m^2/s at 300K
Mannitol	$5{,}7 \times 10^{-10}$
Sucrose	$5{,}2 \times 10^{-10}$
Maltopentaose	$2{,}9 \times 10^{-10}$
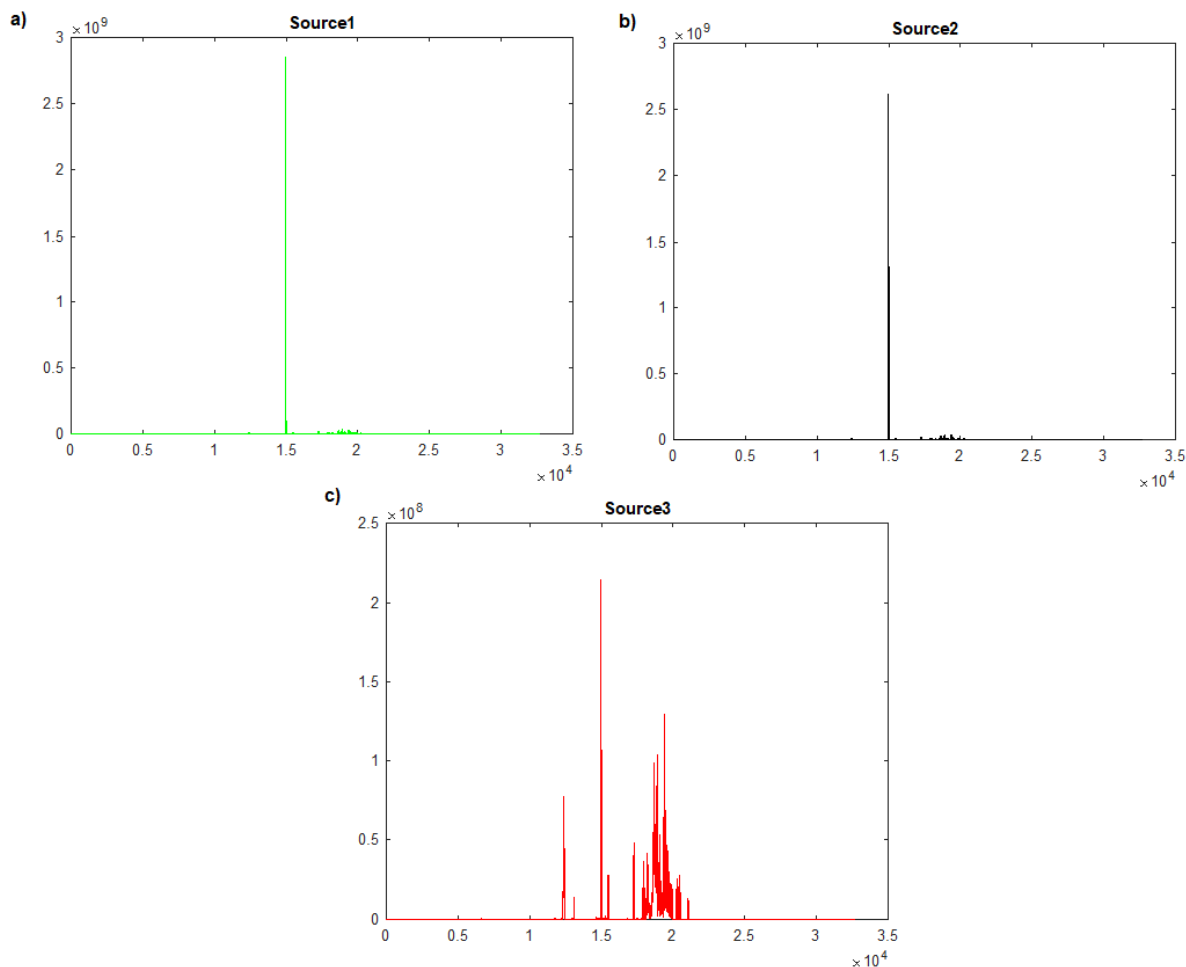
Figure 57 a-c) *Polysaccharide sources not recovered after BSS.*

Again, with this sample, the used algorithms failed to estimate the sources. At this point, we do not have a clear explanation for the failure of the BSS. Our hypothesis is that there are two hurdles which accumulate, the number of sources to find and the too similar diffusion coefficients.

II.4.1.3. Results on micelle sample

The followed experiments were performed on mixtures with a sugar, the sucrose, and a surfactant giving a micellar structure. The idea was to amplify the difference in diffusion coefficients between compounds, always using one sample mixture.

The first sample mixture was obtained using sucrose and sodium dodecyl sulfate, commonly named SDS in D_2O (**Figure 58**).
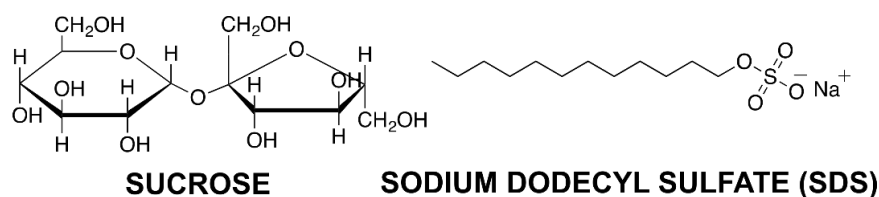
Figure 58 *Sucrose and sodium dodecyl sulfate (SDS) NaC$_{12}$H$_{25}$SO$_4$.*

The critical micellar concentration (CMC) of the SDS is in the range of 8-100 mM at 25°C.[136–139] The concentrations used here to obtain the synthetic mixture were of 30 mM of sucrose and 50 mM of SDS in 600 μL of D_2O.

The application of the BSS was performed on the eleven 1D ^1H DOSY spectra along the 11 points of the gradient strength (**Figure 59**). The initial A_0 and S_0 matrices were built randomly using the random generation function in the MATLAB library. The variations in signal intensities in this unique sample are induced by the gradient strength as an alternative to mime the variation in concentrations.
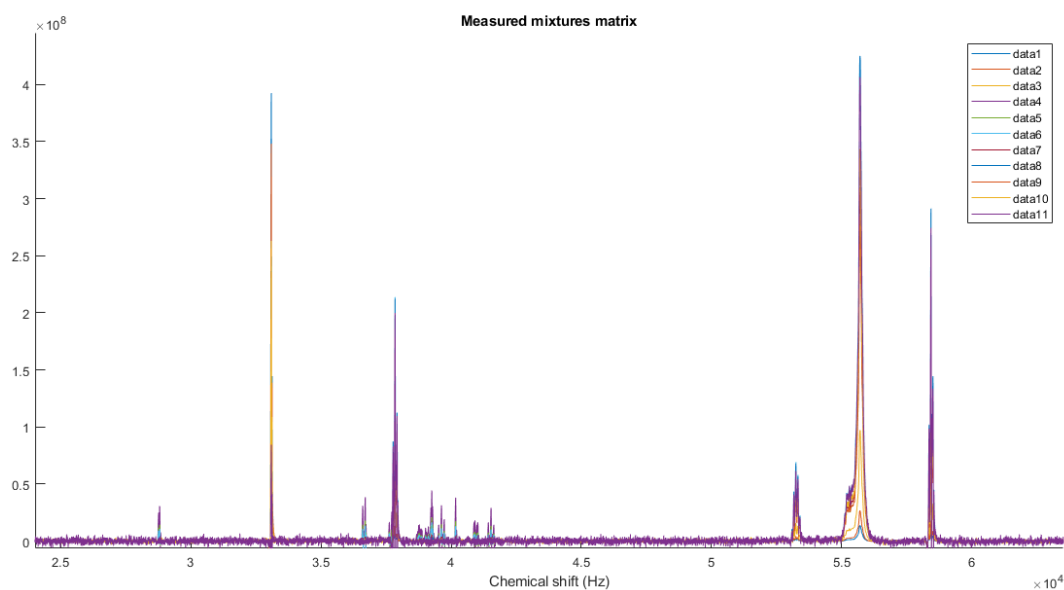
Figure 59 *Eleven superposed 1D proton spectra from the 2D DOSY experiment used for BSS on sucrose and SDS mixture.*

The measured diffusion coefficients are presented in **Table 26**.

The difference from the previous 1D and 2D NMR applications of the BSS is that here the solvent can be considered as a pure source to be estimated since its diffusion is very discriminant. For the previous applications, it could be considered as a source, or a constant peak present in all the other components, or as a region of the spectrum to be suppressed.

Table 26 *Diffusion coefficients and decays for the pure components SDS (green), sucrose (red) and water(water) of the mixture.*

	DIFFUSION COEFFICIENTS m²/s at 300K
SDS	1.32×10^{-10}
Sucrose	4.88×10^{-10}
Water	2.38×10^{-9}
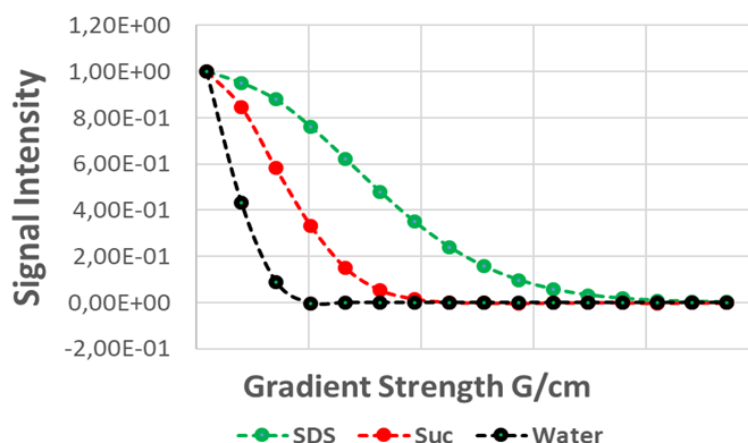
In this sample, the BSS succeeded in extracting the spectra of the three sources, reported in **Figure 60**. The spectrum of the estimated sucrose, in red, contains the residual peaks of the SDS with a low intensity (blue circle in **Figure 60.c**).
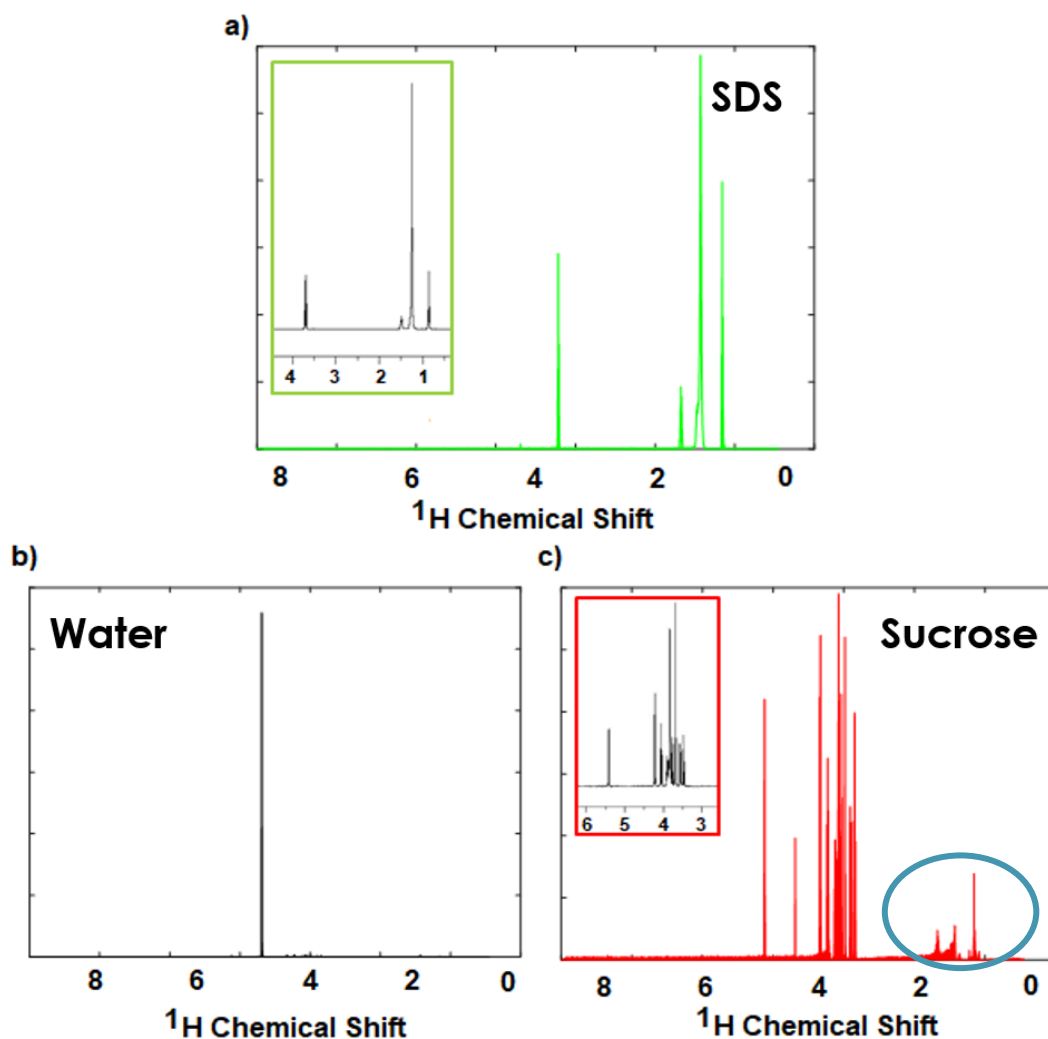
Figure 60 *Estimated spectra of sources. From the top SDS (green), water (black), and sucrose (red). The blue circle indicates the residual peaks of the SDS in the sucrose spectrum with the focus of the reference spectra in the boxes.*

In the following, the objective is to know if the more important difference in the evolution of the intensities of the peaks as a function of the gradients would enable to recover the spectra of the sources without spurious peaks.

The second synthetic sample was composed mixing the sucrose and cetrimonium chloride, named also cetyltrimethylammonium chloride (CTAC), in D_2O (**Figure 61**).
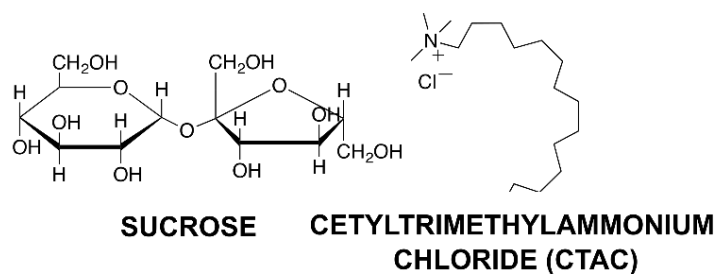
Figure 61 Molecules composing the mixture of sucrose and cetyltrimethylammonium chloride ($C_{19}H_{42}ClN$) (abbreviated as CTAC).

The CMC of CTAC is > 1.3 mM at 25°C.[136] Concentrations used to obtain the synthetic mixture were of 30 mM of sucrose and 50 mM of CTAC in 600 μL of D_2O.

In the 2D DOSY experiment, all the sixteen one-dimensional proton spectra along the diffusion dimension (**Figure 62**) were used to apply the BSS.
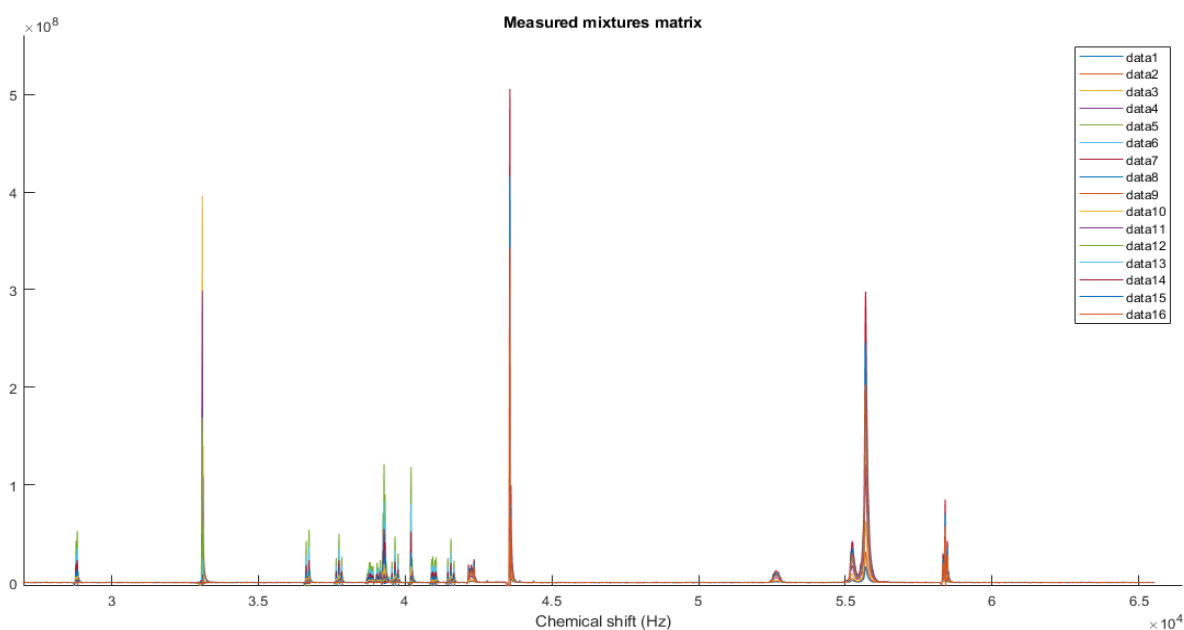
Figure 62 Sixteen superposed 1D spectra of the sample from the 2D DOSY experiment used for BSS on sucrose and CTAC mixture in D_2O.

In this system, the micellar component, the CTAC, and the sucrose have more differentiated translational diffusion decays (**Figure 63**) and diffusion coefficients (

Table 27).
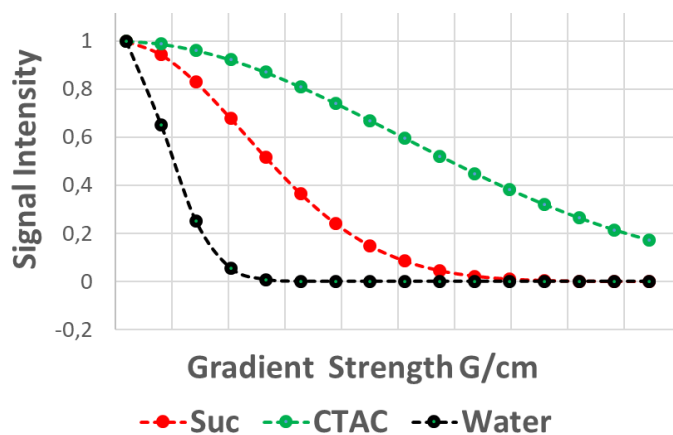
Figure 63   *CTAC diffusion curve with average D = 8.82 x 10^{-11} m^2/s (green), sucrose diffusion curve with D = 4.46 x 10^{-10} m^2/s (red), water diffusion curve with D = 3.31 x 10^{-9} m^2/s (black).*

Table 27 *Diffusion coefficients for the pure components CTAC, sucrose and water of the mixture.*

	DIFFUSION COEFFICIENTS (m²/s) at 300K
CTAC	8.82 x 10^{-11}
Sucrose	4.46 x 10^{-10}
Water	3.31 x 10^{-9}

The estimated source spectra after the BSS are reported in **Figure 64**. In this case, the estimated sucrose spectrum contains less intense residual peaks of CTAC and is less noisy than the previous case with SDS.
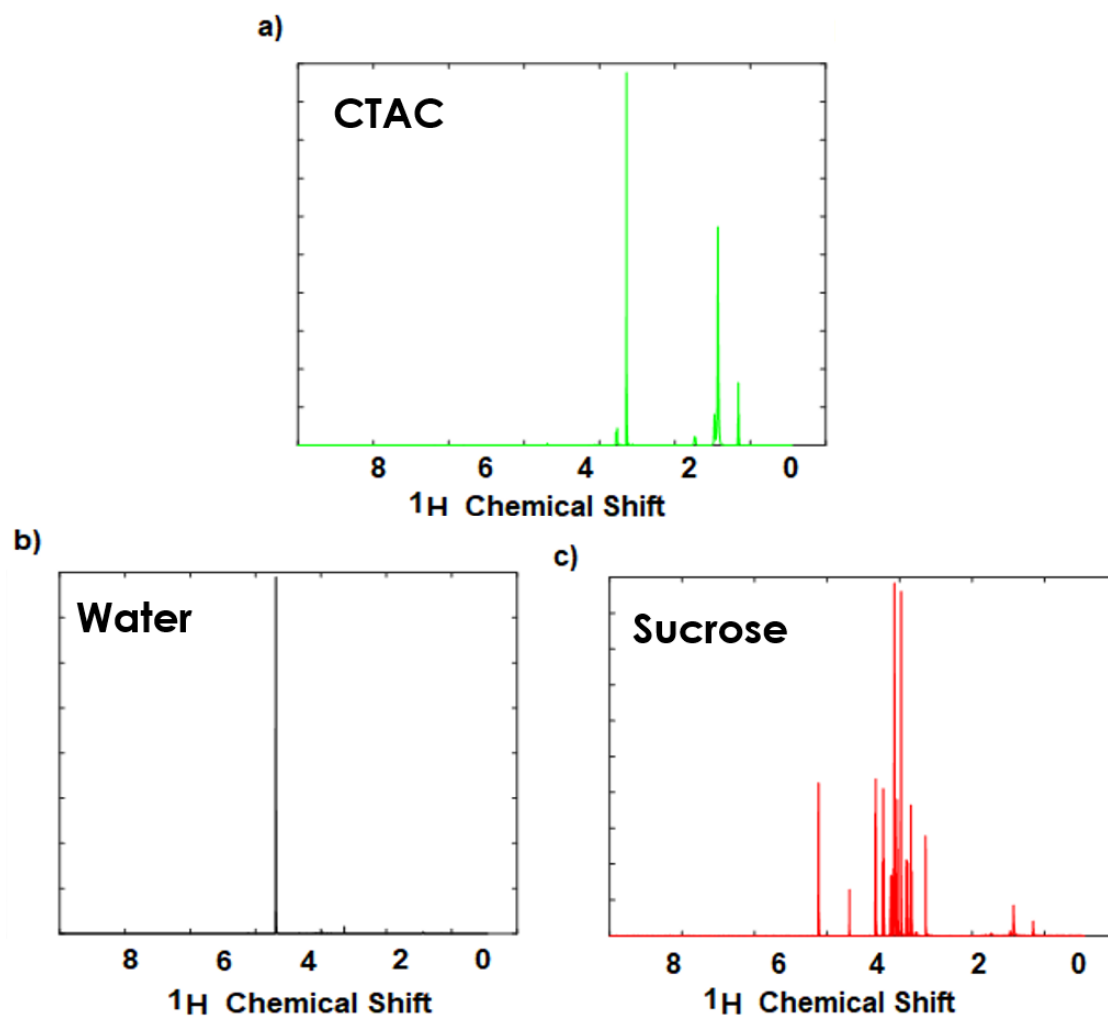
***Figure 64** Estimated spectra of sources. From the top CTAC (green), water (black), and sucrose (red).*

The third synthetic mixture was made up of sucrose and a copolymer in D_2O. Pluronic P123 (P123) is a symmetric triblock copolymer composed of polyethylene oxide at the sides and polypropylene oxide at the centre (**Figure 65**).
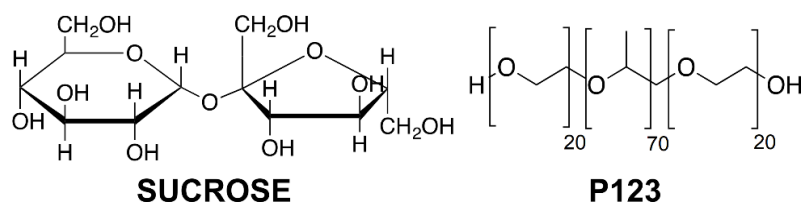
***Figure 65** Molecules composing the mixture of sucrose and P123 in D_2O.*

The characteristic of the central block is the hydrophobic property at above 288 K, and the analyses were carried out at 300 K. Dissolved P123 forms micelles when placed in a selective

solvent such as water and with a concentration above its CMC. The micellization of P123, in fact, has a relatively low CMC in water (0.313 mM at 20°C) and is sensitive to the changes of conditions such as temperature.[140] Concentrations used in this synthetic mixture were of 50 mM for the sucrose and 30 mM for P123 in 600 μL of D_2O.

In **Figure 66**, all the fifteen one-dimensional proton spectra from the 2D DOSY experiment, used as *X* matrix for the BSS, are reported except one outlier.
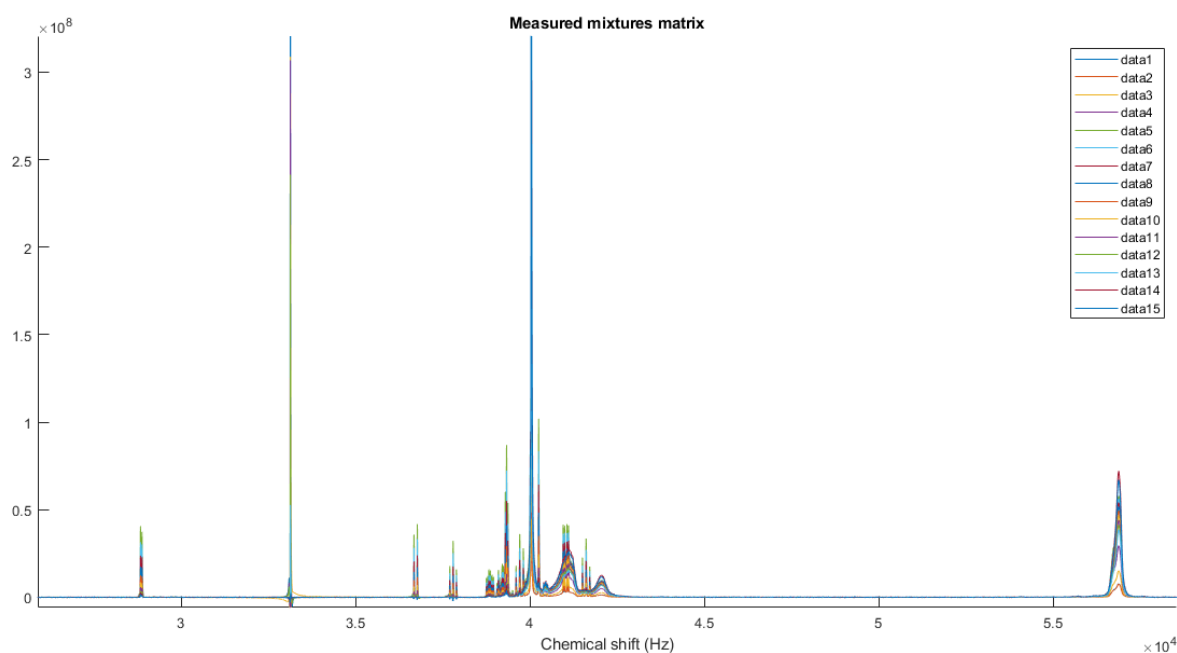
Figure 66 Fifteen 1D proton spectra from the 2D DOSY experiment used for BSS applied on the sucrose and P123 mixture.

The measured diffusion coefficients are presented in **Table 28.**

Table 28 Diffusion coefficients and decays for the pure components P123 (green), sucrose (red) and water(water) of the mixture.

	DIFFUSION COEFFICIENTS (m²/s) at 300K
P123	7.77×10^{-11}
Sucrose	5.25×10^{-10}
Water	2.26×10^{-9}
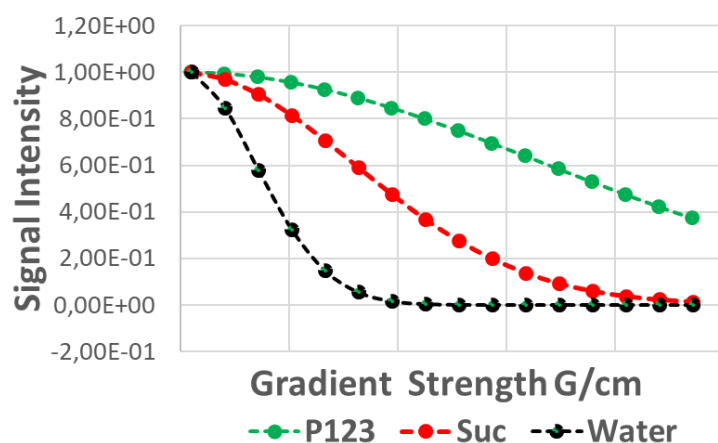
The estimated sources are reported in **Figure 67** with the focus of the reference spectra of each source. An extra peak of the P123 is present in the estimated spectrum of the sucrose, less intense than for the previous mixtures.
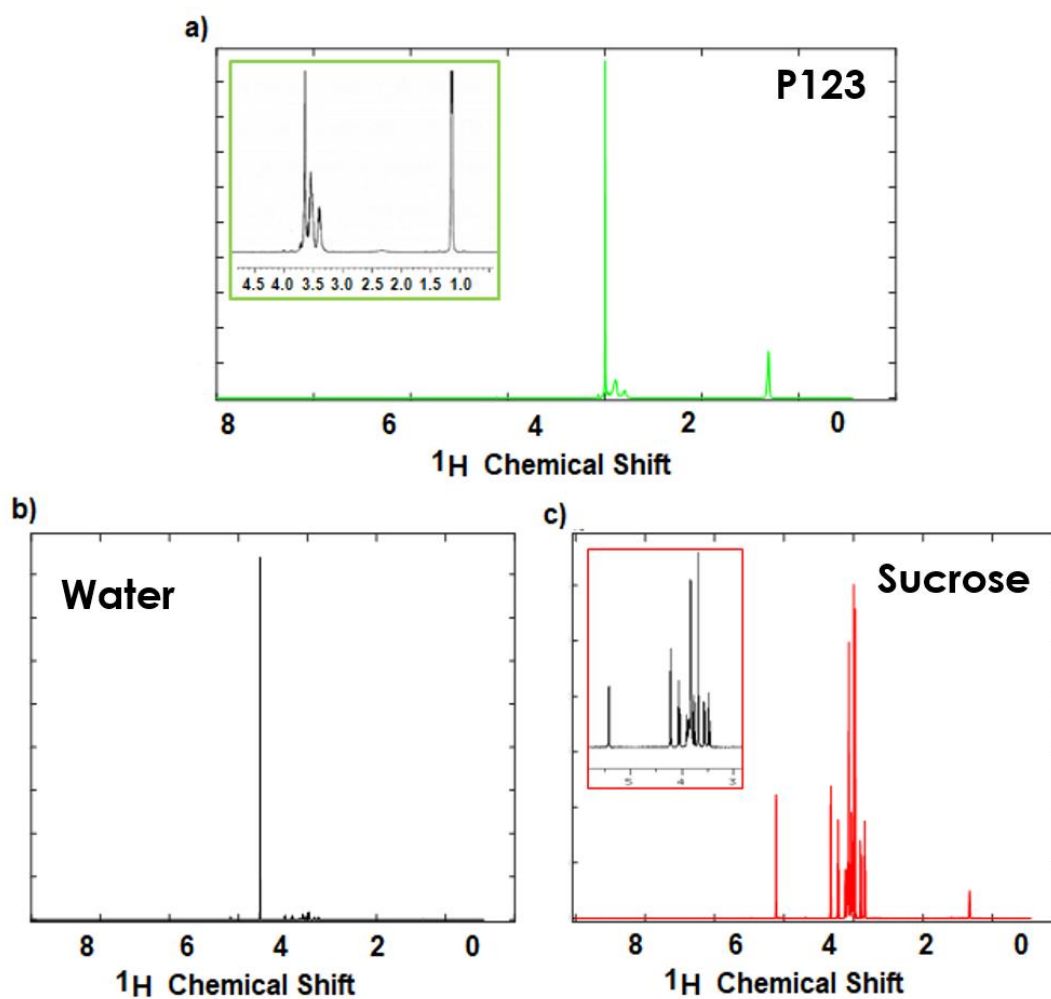
Figure 67 *Estimated spectra of sources. From the top P123 (green), water (black), and sucrose (red).*

II.4.1.4. Conclusions on spectroscopic variations applied to the BSS

This work confirms that, when only one sample is available, it is possible to introduce and exploit spectroscopic perturbations that mimic the concertation variations to blindly decompose NMR spectra, without a prior known of intensity evolution.

Unlike the founding work of Nuzillard and co-workers,[133] the samples used here were composed of at least three molecules. Despite relatively different diffusion coefficients, especially for polysaccharides, the algorithms we used did not succeed in decomposing the mixtures. We think that the number of sources to decompose is a limitation, at this level, that can add up to the non-sufficient difference in the diffusion coefficients. For micelle samples, with significant diffusion differences (an order of magnitude), the algorithms were efficient in estimating the sources spectra, despite a few spurious peaks.

It should be noted that the BSS succeeded in finding the sources composing the mixtures as long as the number of spectra m in the matrix X was greater than or equal to the number of sources to be estimated k ($m \geq k$). However, the more the number of m spectra increased, the better the estimate.

Further methods to introduce other spectroscopic perturbations to blindly decompose the mixtures can be envisaged. One can imagine an approach using selective saturation in mixtures in which the spin diffusion is present due to the size of the molecules or by the addition of viscous solvent as proposed by the group of J.M. Nuzillard.[141–143]

III. Application of NMR developments on real samples

III.1. Assessment of $T_{1\rho}$ filter in metabolomics

The first part of this section presents a study that investigates the use of the rotating-frame relaxation $T_{1\rho}$ filter in biological samples in the context of metabolomics. This work is carried out both in solution and HR-MAS NMR. It draws detailed comparisons with conventional CPMG filtering as well as the more recent approach PROJECT based on the perfect echo approach for T_2 filtering.[44] The data showed that the $T_{1\rho}$ filter is suitable for metabolomic investigation involving multivariate statistical analysis on sample cohorts, with results equivalent to the ones with T_2 filter.

III.1.1 General Principles of Metabolomics

Metabolomics is a recent versatile science that was introduced in the 2000s following genomics and proteomics of which it is often complementary. Its purpose is to study the dynamic and global metabolic response of a biological system following biological stimuli such as disease or treatment, or due to environmental or genetic perturbations (*e.g.* drug, diet, lifestyle, environment, stimuli, genetic and other modulations) by the statistical analysis of biological data obtained by spectroscopic analysis, such as NMR or Mass Spectrometry (MS). The objective is to characterize and identify the metabolites produced, used, and excreted by cells, organs, or organisms, to identify biomarkers associated with well-defined systemic states in order to highlight the metabolic involved pathways.[144] [145] [146] [147]

The metabolites are usually defined as any small molecule with a molecular weight less than 1.5 kDa including several organic species such as amino acids, fatty acids, carbohydrates, vitamins, but they can also contain inorganic species.[148] [149] Metabolome, by analogy with the

notion of genome and proteome, is defined as a collection of all metabolites, endogenous or exogenous, which can be detected in cells, tissues, extracts or bio fluids.

The main strategy adopted is to obtain a classification based on the statistical analysis of the metabolic profiles mainly obtained by NMR and /or MS. Thus, the information that is extracted and exploited concerns the modulations of metabolite concentrations induced by the state of the sample. In practice, the experimental metabolomic approach can be described in three main steps (**Figure 68**). Firstly, the simultaneous identification and quantification of metabolites are achieved by analytical techniques. The obtained raw data are then subjected to multivariate statistical analysis. At the end, the metabolic signature is examined. Metabolite characteristics are identified from spectral data, a biological interpretation of experimental data is performed, and research for individual or composite biomarkers is also performed.
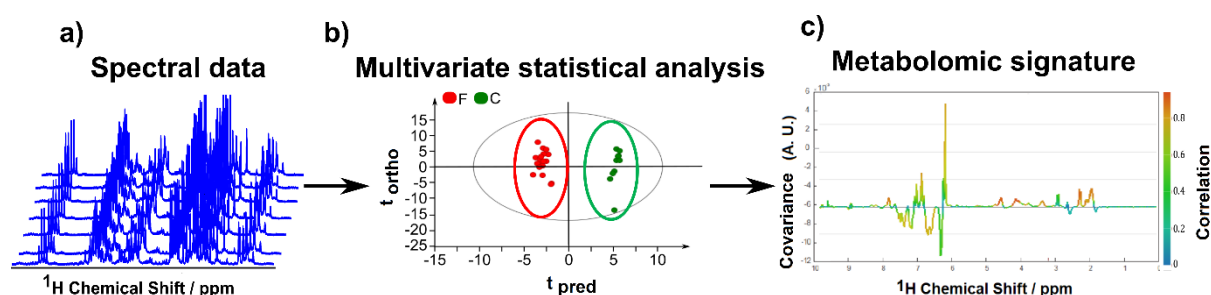
Figure 68 *Strategy for metabolomic analysis:* *a)*Acquisition of spectral data, *b)* Classification of samples by multivariate statistical analysis, *c)* Identification of metabolic signature (metabolic phenotype).

III.1.2 NMR Relaxation Filters in Metabolomics

The number of samples submitted to high-throughput analysis without complete prior components separation and with minimal pretreatment (protein removal in blood, for example) has been increased by the development of NMR-based metabolomics.[35,129–131,146,149,150] To face the obvious complexity of NMR spectra of biological samples, such as blood plasma and tissues, containing different molecules of widely varying structures and sizes (metabolites, proteins, lipoproteins, lipids), several spectral editing NMR sequences have been proposed to

suppress the signal of macromolecules which are not usually taken into account in metabolomics (proteins, lipoproteins and lipids).[151] In general, the complete elimination or the attenuation of macromolecule signals exploits the large difference in transverse or spin–spin relaxation T_2 times between small molecules and macromolecules using relaxation-weighted experiments to filter out the NMR signals from proteins and other biomolecules. In the established metabolomic protocols, the only filter used is the T_2-based one with the CPMG pulse sequence (**Figure 15**).[41,152] As presented before (paragraph I.1.2.2.1), the relaxation in the transverse plan can be characterized by two relaxation times, T_2 and $T_{1\rho}$. Even if the relaxation mechanisms are not the same,[7] the same macroscopic phenomenon prevails with the same equation describing the signal evolution as a function of time (**Equation 1.4** and **1.5**):

A major drawback of the T_2 based filter with the conventional CPMG pulse sequence is the echo modulation arising from homonuclear scalar coupling which is not completely refocused by the multiple CPMG block even when the inter-pulse frequency ($1/\tau$, where τ is the inter-pulse delay) is much larger than J coupling values ($1/\tau >>> J$).[152,153] Several strategies have been proposed to overcome these issues.[77,154–158] However, the proposed approaches so far are limited to scalar coupled two-spin systems and do not allow a broad-based application. More recently, *Aguilar* and co-workers[44] formally demonstrated that the effect of the so-called "perfect echo" sequence, proposed by Takegoshi *et al.*,[43] is not limited to AX spin systems but can be extended to multiple coupled spins. Hence, CPMG with a 90°y refocusing pulse at the midpoint of a double spin-echo was introduced as PROJECT and was successfully used in metabolomics, particularly for HR-MAS studies on tissue samples, where long inter-pulse delays must be used for rotor synchronization.[159–161] **Figure 69** presents a standard ^{1}H NMR spectrum of liver sample (**Figure 69.a**) and a T_2-filtred spectrum using the standard CPMG NMR spin-echo (**Figure 69.b**), both with water presaturation pulse during the relaxation delay.
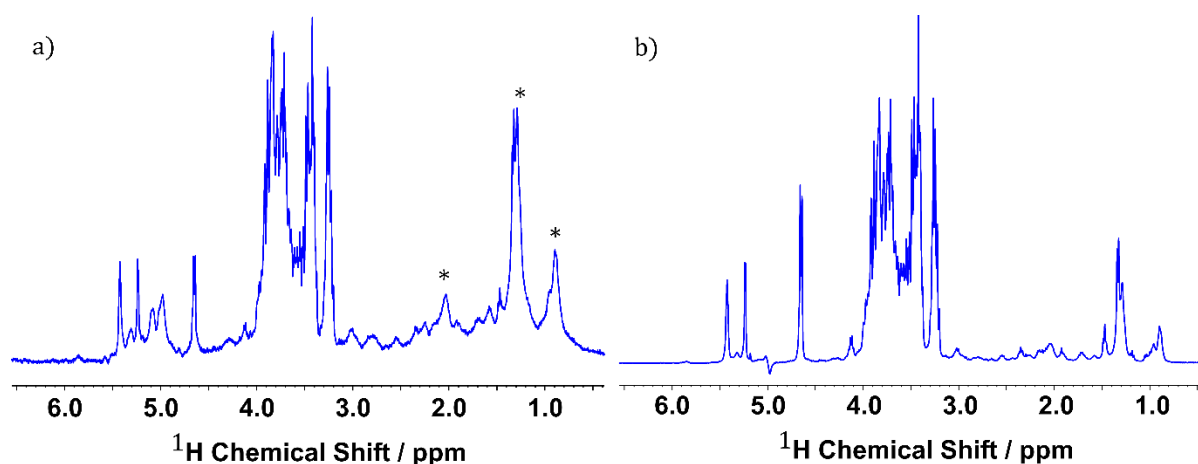
Figure 69 a) *A zgpr ^{1}H spectrum on the left and **b)** a relaxation-edited CPMGpr ^{1}H spectrum of rat liver at 400 MHz HR-MAS. The asterisks * indicate the fatty acids which are attenuated in CPMG experiment as confirmed in the literature.[162]*

When dealing with a proton spectrum involving many multiplets, a simple alternative can be used, the relaxation time in the rotating frame, $T_{1\rho}$. It is measured from the decay of nuclear magnetization locked in the transverse plan by an *rf* field by using the spin-lock as described in **Figure 17**.[163] An immediate advantage of this relaxation time is that there is no precession during the evolution period of the magnetization. This results in the absence of modulations related to possible scalar *J* couplings.[159] Although similar results are obtained using T_2 or $T_{1\rho}$ filters, and despite the advantages of $T_{1\rho}$, there is no application in mixture analysis and it has never been used in metabolomics.[164,165]

Generally, the CPMG filter duration used in metabolomics is between 60 ms and 120 ms depending on the studied samples (serum, plasma, different types of tissues, *etc.*) which is sufficient to attenuate the resonances from large macromolecules.

III.1.3 Samples used for this study

The collection of the samples used for this study is detailed in our previous publications by Tranchida *et al.*[166,167] Serum and liver samples were collected from rats following two diets: (i)

a standard diet (control group – *C*) and (ii) a High Fructose and Saturated fatty acids diet (*HFS* group). The distribution of the groups is given in **Table 29**.

Table 29 Number of independent samples in each analysed group (control and HFS) of rat serum and liver.

	Control group	HFS group
Serum	6	12
Liver	6	12

Serum samples were prepared adding 200 μl of serum with 400 μl of 0.9 g·L^{-1} saline solution in D$_2$O. For the HR-MAS study, about 15 mg of liver tissue were placed into a 30 μl cylindrical disposable insert where 10 μl of D$_2$O were added. The insert was then placed into an 80 μl 4 mm ZrO$_2$ HR-MAS rotor.

III.1.4 NMR recording conditions

Three pulse sequences were implemented: (i) a standard CPMG NMR spin-echo sequence from the Bruker pulse program library, preceded by a water presaturation pulse during the relaxation delay ([presat $- 90° - (\tau - 180° - \tau)_n$]) (**Figure 15**); (ii) a standard $T_{1\rho}$ sequence, also preceded by a water presaturation pulse during the relaxation delay (presat $- 90\phi° - (SL)_{\phi\pm90°}$]) (the pulse sequence and implementation details are given in **Figure 17**); and (iii) the presaturated PROJECT sequence ([presat $- 90° - (\tau - 180° - \tau - 90° - \tau - 180° - \tau -)_n$]) from Aguilar et al.,[44] (**Figure 16**) used for HR-MAS experiments.

For liquid-state NMR, the experiments were performed at 293 K. For HR-MAS NMR Spectroscopy, all spectra were acquired at 277 K. For CPMG and PROJECT, the inter-pulse delay had to be synchronized with the two MAS periods used in this study, $\tau = 1$ ms and 250 μs for $\nu_{rotation} = 1000$ and 4000 Hz, respectively.[159,167] The filter length was 100 ms. To reach this length, it is necessary to take into account the number of inter-pulse delays inside the loops

of the two sequences, two for CPMG and four for PROJECT. The distribution of these parameters is summarized in **Table 30.**

The same filter length was used for the $T_{1\rho}$ spectra without any calculation by adjusting the spin-lock pulse, "p3" in **Figure 17**. For each experiment, the spectral width was set to 4500 Hz and sampled with 32,768 data points, leading to an acquisition time of 2.26 s and a total experiment time of 10 min.

Table 30 Parameters τ (inter-pulse delays) and n (number of loops) used for CPMG and PROJECT to obtain a filter length of 100 ms for two MAS rates (1000 and 4000 Hz).

$\nu_{Rotation}$	CPMG (τ-180°-τ)n	PROJECT (τ-180°-τ-90- τ-180- τ-)n
1000 Hz	τ = 1 m / n = 50	τ = 1 m / n = 25
4000 Hz	τ = 250 us / n = 200	τ = 250 us / n = 100

III.1.5 Statistical Analysis

For metabolomics analysis, the most widely used NMR experiment yields a one-dimensional ^{1}H spectrum containing complex patterns which constitute a metabolic fingerprint describing all visible metabolites at a given time. Statistical analyses allow providing understandable and usable information about these metabolic fingerprints. The proton NMR spectra are divided into 1000 to 10000 buckets characterized by a width and an integral proportional to the concentration of the metabolites. This first step facilitates pattern identification, or group clustering, based on spectral pattern differences. In general, two major types of pattern recognition processes are used, unsupervised and supervised. Unsupervised data analysis, such as Principal Component Analysis (PCA), already introduced before, measures the intrinsic variation in data sets, whereas the supervised approach, such as Orthogonalized Projections to Latent Structures – Discriminant Analysis (OPLS-DA), uses prior information (the group distribution for example) to generate the clusters of patterns. The comprehensive representations of the data that allow to visualize the discrimination between

groups are called "score plot" (**Figure 68.b**). In addition, the "loading plot" representation (**Figure 68.c**) allows highlighting specific spectral regions which are responsible for group clustering and linking them to specific metabolites according to their NMR chemical shifts.

The ^{1}H NMR spectra were exported to *AMIX* software and divided into 0.001 ppm-width buckets. To remove the effects of possible variations in the water suppression efficiency, the region of the water signal between 4.7 and 5.2 ppm was suppressed. The obtained data set was then normalized to the total spectrum intensity and scaled to Unit-Variance (UV) and imported to SIMCA P-14 software (Umetrics, Umeå, Sweden) for the multivariate analysis.

OPLS-DA was then performed on the data set to discriminate between C and HFS groups.[168] The robustness of each discriminant model was then validated using the goodness-of-fit, R^2_Y, and the predictive parameter, Q^2_y, that were calculated. The highest the Q^2, the more the model can be considered predictive. In model systems, Q^2 are high typically > 0.7 or 0.8 for predictive models. For biological samples, it is different. The inter-individual variability among biological samples is quite regular and therefore the overall changes are weak between different groups, therefore we considered that a predictive model is statistically robust for Q^2_y value ≥ 0.4.[169,170]

III.1.6 Impact of $T_{1\rho}$ filter on sample heating

Some studies involving the $T_{1\rho}$ filter evoke a problem of sample heating during the application of an excessive and long spin-lock.[171] This effect is inherent to the nature of *rf* pulse and due to its electric component. We evaluated this using a sample of methanol as a molecular thermometer to measure the actual temperature before and after the application of a 480 ms spin-lock of different amplitudes, *i.e.* 2000 Hz, 4000 Hz, and 5000 Hz. For each amplitude, three spectra were recorded using a one pulse sequence ("zg" according to Bruker nomenclature), after 1, 8, 64 and 128 dummy scans. The results are reported in **Table 31**.

Table 31 *Temperatures measured from ^{1}H spectra of methanol-d4 using one pulse (zg), CPMG and $T_{1\rho}$ sequences with the same filter duration (480 ms), different number of dummy scans (DSs) were used, with number of scans NS=1. The target sample temperature was set to 277 K. The difference between the target and actual temperature is reported.*

Sequence used	Number of dummy scans	Spin-lock amplitude (Hz)	Δ Temperature (K)
zg			=0
CPMG			+0.10
$T_{1\rho}$	1	2000	+0.16
		4000	+0.12
		5000	+0.15
zg			=0
CPMG			+0.10
$T_{1\rho}$	7	2000	+0.07
		4000	+0.14
		5000	+0.15
zg			+0.02
CPMG	63		+0.13
$T_{1\rho}$		2000	+0.12
		4000	+0.10
		5000	+0.17
zg			+0.02
CPMG			+0.07
$T_{1\rho}$	128	2000	+0.08
		4000	+0.15
		5000	+0.19

Recorded temperatures show the same small increase as for the CPMG pulse sequence, less than 0.2°C, regardless of the number of used dummy scans. The use of $T_{1\rho}$ for these amplitudes and durations remains safe since it does not cause sample overheating.

III.1.7 Spin-lock amplitude

Conventionally the spin-lock pulse is obtained by a continuous *rf* wave with an appropriate amplitude which avoids the shift between the carrier frequency and the resonance frequencies ("offset effect") and unsuitable power deposition due to excessive amplitude.

$T_{1\rho}$ general pulse sequence is reported in **Figure 17**. For example, the resulting relaxation-edited spectrum is reported in **Figure 70** when using low ν_1 for spin-lock amplitude such as 1000 Hz and putting the carrier ν on the water peak at 4.8 ppm, there is a low spectral coverage and an offset effect is visible around 1 ppm.
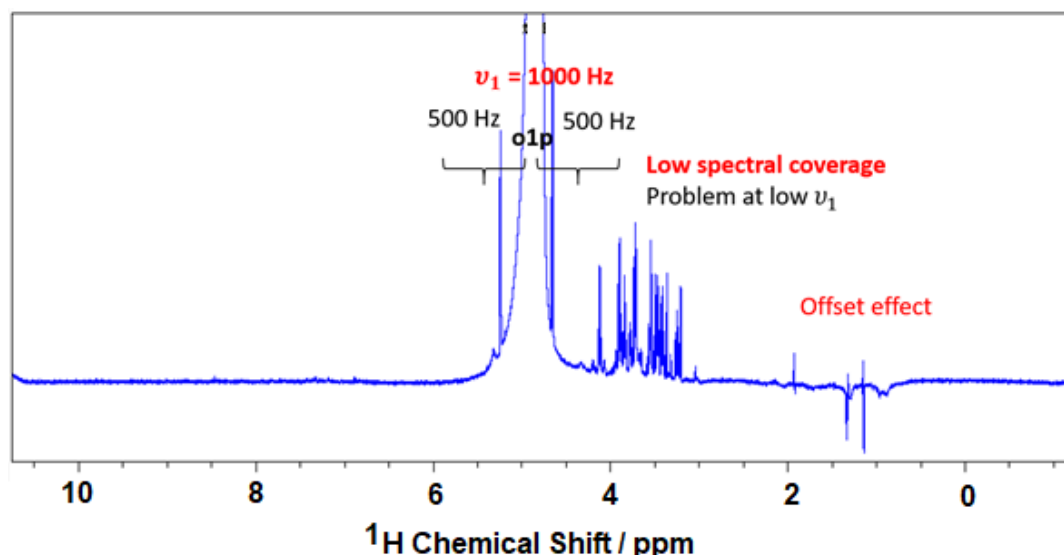
Figure 70 *Relaxation-edited $T_{1\rho}$ spectrum of proton using only one channel with the carrier frequency for a pulse v_1=1000 Hz centered on the water peak. At low v_1, due to the low spectral coverage, an offset effect is present.*

Five spectra of serum were recorded with CPMG filter using two inter-pulse delays (200 μs and 600 μs) and $T_{1\rho}$ filter with different spin-lock intensities expressed as $\frac{\gamma B_1}{2\pi}$ (2 kHz, 4 kHz, and 5 kHz). **Figure 71** represents the 480 ms T_2- and $T_{1\rho}$-edited proton spectra with the carrier ν at 4.7 ppm (water frequency, also used for presaturation). The first finding was that the same filtration effect was retrieved whatever the sequence used (**Figure 71.a**). However, the three $T_{1\rho}$ spectra obtained with different B_1 showed expected differences due to the offset effect.[172] Around the carrier frequency, the integrals of the signals remained unchanged when B_1 amplitude increased (**Figure 71.b**), while the integrals of the peaks at the edges of the spectra increased (**Figure 71.c**). Moreover, we observed a decrease of the resolution when increasing B_1 (therefore of the nutation frequency, $v_1 = \frac{\gamma B_1}{2\pi}$) leading to a broadening of the width at half-height (*e.g.* the creatine singlet at 3.03 ppm) (**Figure 71.a**).
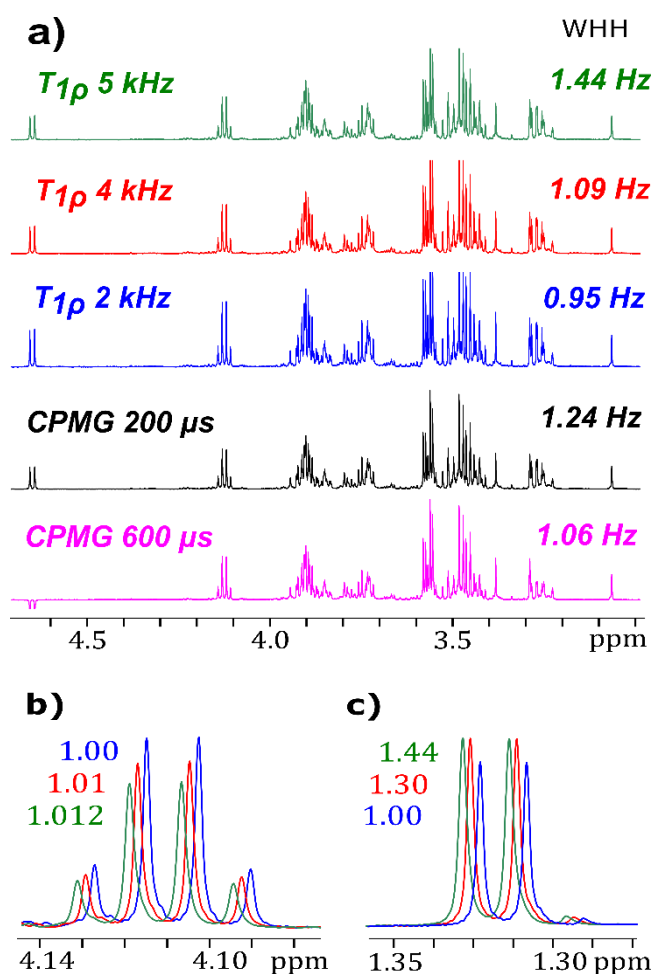
Figure 71 (a) $T_{1\rho}$ *and* T_2 *filtered spectra of serum recorded at different spin-lock amplitudes, from top to bottom: 5 kHz, 4 kHz, and 2 kHz, and CPMG with inter-pulse delays of 200 and 600 µs. Quartet (b) and doublet (c) of lactate acquired at 5 kHz (green), 4 kHz (red) and 4 kHz (blue). The relative integrals are reported for each peak with the same colour. The widths at half height (WHH) of the peak at 3.03 ppm (creatine singlet) are reported on the right of each spectrum.*

This trend was identical for the CPMG sequence, the resolution obtained with an inter-pulse delay of 600 µs, *i.e.* a CPMG pulse frequency ($\nu_{CPMG} = \frac{1}{4\tau}$, with τ the inter-pulse delay of the CPMG block) of 416.66 Hz, was higher than for the one obtained with an inter-pulse delay of 200 µs, *i.e.* a CPMG pulse frequency of 1250 Hz. This effect is summarized in **Table 32**.

Table 32 Width at half height of the creatine singlet (3.03 ppm) [Average over five randomly selected samples].

	$CPMG_{600\mu s}^{416Hz}$	$CPMG_{200\mu s}^{1.25\,kHz}$	$T_{1\rho}^{2kHz}$	$T_{1\rho}^{4kHz}$	$T_{1\rho}^{5kHz}$
WHH	1.06 Hz	1.24 Hz	0.95 Hz	1.09 Hz	1.44 Hz

The best resolution is obtained with $T_{1\rho}^{2kHz}$ but it is obvious that, with this power, the spectral coverage is not optimal to avoid offset effects. With the $T_{1\rho}^{5kHz}$, two disadvantages add up: i) the power that could be excessive, and ii) the decrease of the resolution. The $T_{1\rho}^{4kHz}$ seems to be, for our study, the better amplitude allowing to have an equal resolution of the $CPMG_{600\mu s}^{416\,Hz}$ sequence (see **Table 32**). However, the latter cannot be used since the inter-pulse delay does not allow the refocusing of the modulations due to scalar couplings as shown in **Figure 72** with the negative doublet at 4.65 ppm. In liquid-state metabolomic studies by NMR, with the commonly used CPMG filter ($CPMG_{200\mu s}^{1,25kHz}$), this effect is limited and invisible.

The impact of the offset effect on the intensities at the edge of the spectra can be explained qualitatively by using the vector model (**Figure 72**). When the nuclei are on resonance ($v_0=v_r$), they perceive an effective field equivalent to the field B_1 but if the shift between the carrier frequency and the resonant frequencies becomes important, the effective field in the rotating frame, B_{eff}, is defined as the vector sum between the vector of the field B_1 (aligned along the axis of the spin-lock, y for example) and the vector of the residual field B_{res} aligned along the z- axis of the static magnetic field B_0 (**Figure 72**). Knowing that B_{res} is proportional to the offset (v_0-v_r) and given by:

$$B_{res} = 2\pi \frac{v_0 - v_r}{\gamma} = B_0 - \frac{2\pi v_r}{\gamma} \tag{3.1}$$

In this case, the spin-lock axis is not aligned along y but along a shifted axis with an angle θ which increases when the shift between the resonance frequency v_r and the carrier frequency v_0 increases (an increase of the B_{res}, **Equation 3.1**), and obviously when the B_1 field decreases (B_1' in **Figure 72**). The magnetization initially aligned along the y-axis (before the spin-lock) will be aligned along the B_{eff}, then appear components perpendicular to the B_{eff} axis which will be destroyed (defocused) and a collinear component to B_{eff} which will be locked and therefore

maintained. In addition, the amplitude of the effective field becomes greater than that of B_1 due to the vector sum, $B_{eff} = \sqrt{B_1^2 + B_{res}^2}$, so the precession will be greater around B_{eff}.
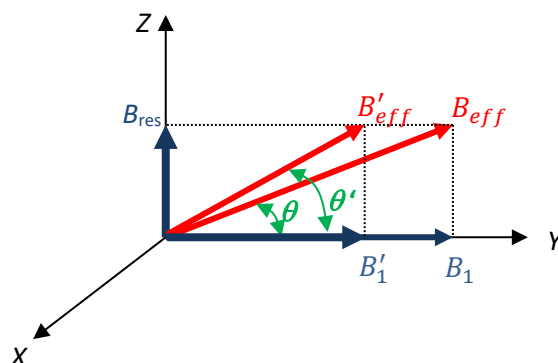
Figure 72 *Representation of the magnetic fields present during spinlock application. In the rotating frame, the effective field Beff is the vector sum of the residual field Bres and the B1 field. The tilt angle, ϑ, is defined as the angle between Beff and Bres.*

For the same offset (same B_{res}, in **Figure 72**), the increase of the amplitude of the B_1 field leads to the decrease of the angle θ. In this case, the component perpendicular to B_{eff} (at the origin of the loss of intensity) will be less important, which explains the increase of the signal at the end of the spectra when increasing the spin-lock amplitude.

In other words, to compensate for the difference between the carrier frequency and the resonant frequencies (B_{res}), it is necessary to systematically use a spin-lock amplitude that covers all the spectral width (the entire spectrum).

III.1.8 $T_{1\rho}$ Measurements for Some Chosen Metabolites

The measurements of the relaxation times $T_{1\rho}$ and T_2 for some chosen metabolites in serum confirm that these two relaxation times are substantially similar. To avoid the offset effect, a sequence using two channels with two carrier frequencies was used (**Figure 73**), the first for water presaturation and the second placed at 2,2 ppm centered on the peaks of interest.

Above 3000 Hz, we observed that the measured of $T_{1\rho}$ were insensitive to the position of the

carrier frequency when placed on the water peak.
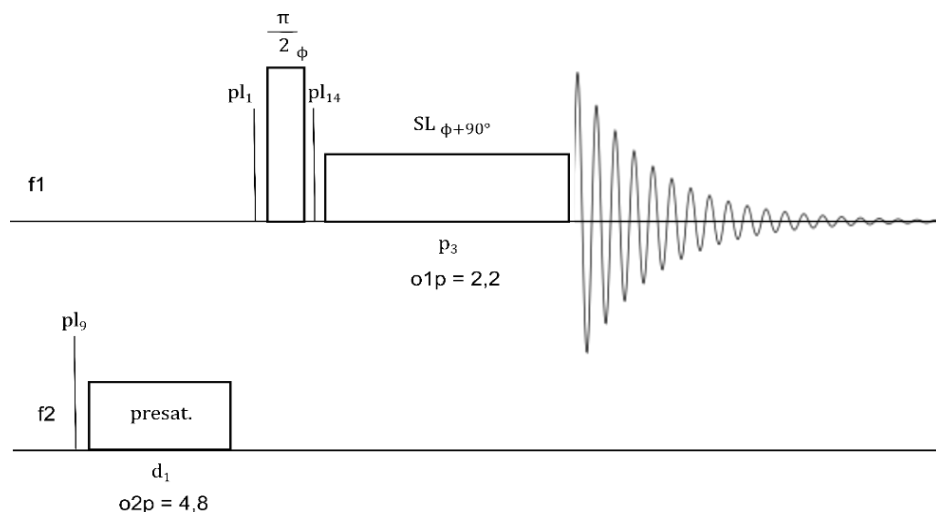
Figure 73 *The best edited-relaxation spectra come from the T_{1p} pulse sequence using two channels, F_1 and F_2. The second channel F_2 is used for the pulse on the water frequency (o2p=4,8 ppm). The first channel F_1 is for the hard pulse to let the magnetization going on the transverse plane, and the carrier frequency for the spin-lock wave set on the center of the region with the peaks of interest, at 2.2 ppm (o1p), between the water at 4.8 ppm and 0 ppm, for a good spectral coverage.*

Figure 74 depicts the evolution of the relaxation rates (inverse of the relaxation times) $R_{1\rho}$ as a

function of the nutation frequency (ν_1). The stability of the $R_{1\rho}$ measurements (**Figure 74.a, b**)

depends largely on the slow molecular dynamics and the offset effect. These effects are visible

at low spin-lock amplitudes, below $\nu_1 = 2.5$ kHz. However, the dispersion curves show that

above 3 kHz, $R_{1\rho}$ stabilizes for most peaks. These results confirm that the previously designated

spin-lock amplitude (4 kHz, red circle) is an adequate value for serum samples.
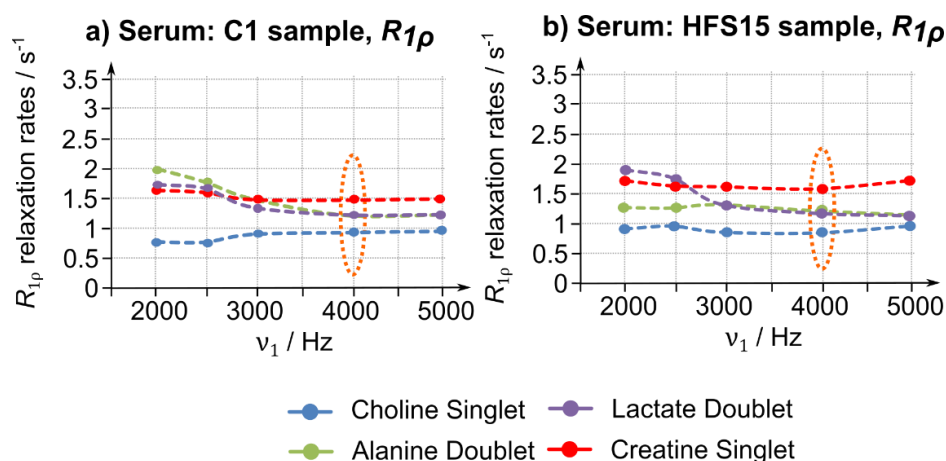
Figure 74 $R_{1\rho}$ *relaxation rates as a function of υ_1, for C and HFS serum samples relative to the peaks of four metabolites. (**a**) $R_{1\rho}$ for the C1 sample and (**b**) $R_{1\rho}$ for the HFS15 sample. Dotted lines are used to help visually track the progress of $R_{1\rho}$.*

III.1.9 Effect of B₁ Amplitudes on Statistical Analysis

To compare the separation capability of $T_{1\rho}$ and T_2 filtered spectra, OPLS-DA was performed to discriminate the C and HFS groups. The score plots are presented in **Figure 75**, where, the two OPLS component models (1 predictive + 1 orthogonal component) display a clear discrimination between HFS and Control samples with comparable statistical results obtained for all sequences (**Table 33**).

Table 33 *The OPLS-DA R^2_Y, Q^2_Y and p-values for HFS vs. C serum samples.*

A	N = 18	R^2_Y	Q^2_Y	p-value
1+1+0	CPMG	0,887	0,678	0,0034
1+1+0	$T_{1\rho}^{2kHz}$	0,898	0787	0,00027
1+1+0	$T_{1\rho}^{4kHz}$	0,854	0,645	0,0062
1+1+0	$T_{1\rho}^{5kHz}$	0,886	0,737	0,00098

These results show that the properties of each of the experiments used are reproducible and allow discrimination between groups by multivariate analysis (**Figure 75**). In addition, the loading plots exhibit the same profile for all the sequences with identical correlation coefficients.
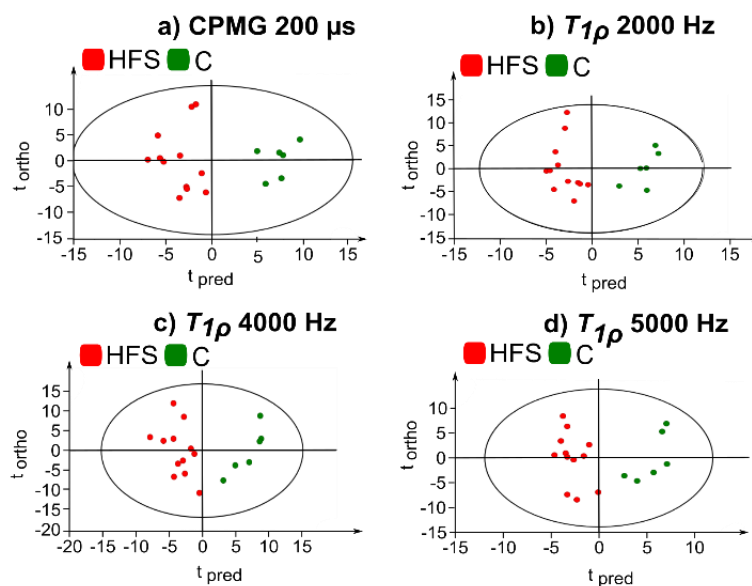
Figure 75 *OPLS-DA score plots for N=18 rat serum samples comparing the distribution of the HFS group with the C group using different sequences: (a) CPMG 200 μs, (b) $T_{1\rho}$ at 2000 Hz, (c) $T_{1\rho}$ at 4000 Hz, (d) $T_{1\rho}$ at 5000 Hz.*

The choice of the appropriate experiment may be guided by reviewing the pros and cons of each sequence. The $T_{1\rho}^{4kHz}$ seems to be the best compromise between optimal spectral coverage and the use of moderated spin-lock amplitude. In this study, the filter duration of 480 ms was applied and chosen by eye to eliminate all the broad signals, including those of lipids, to process only the signals of the metabolites. Even if we consider that the amplitude of $T_{1\rho}^{5kHz}$ is not excessive, especially with the filter lengths commonly used in metabolomics (between 60 ms and 100 ms), we must point out the only drawback which concerns the lower resolution obtained with this amplitude, which remains however safe to use. $T_{1\rho}^{2kHz}$, as seen above, is characterized by a significant offset effect and should be ruled out.

III.1.10 Assessment of $T_{1\rho}$ filter on HR-MAS Metabolomic Studies

The relevance of the use of the $T_{1\rho}$ filter for HR-MAS metabolomics has never been discussed in the literature. In this section, we propose to assess this relaxation time on rat liver samples.

It should be noted that in HR-MAS studies of tissues, the possible degradation of the samples is an important issue. For a relatively high spinning rate (*e.g.*, 4 kHz), tissue degradation can lead to the release of metabolites into the extracellular medium, resulting in narrowing of the peaks over the experiments and therefore bias in the comparison. However, the impact of freezing and prolonged HR-MAS analysis on the metabolic profile was assessed and shows that after 1.5 h, with a spinning rate of 5000 Hz, the levels of certain metabolites start to change.[173] In this section, the relaxation measurement was performed on freshly prepared samples from pieces of liver, stored at −80 °C, with a total experiment time of about 1 hour 40 min (10 min per $T_{1\rho}$ measurement), during which we did not observe significant changes in the spectra (**Figure 76**). For the statistical assessment, the total experiment time was about 1 hour (10 min per experiment, CPMG / PROJECT / $T_{1\rho}$), two MAS rates were used each time, starting with the slow spinning rate (1000 Hz). Moreover, tissue degradation during the NMR experiments was avoided by maintaining the sample at 4 °C with the Bruker cooling unit (BCU).
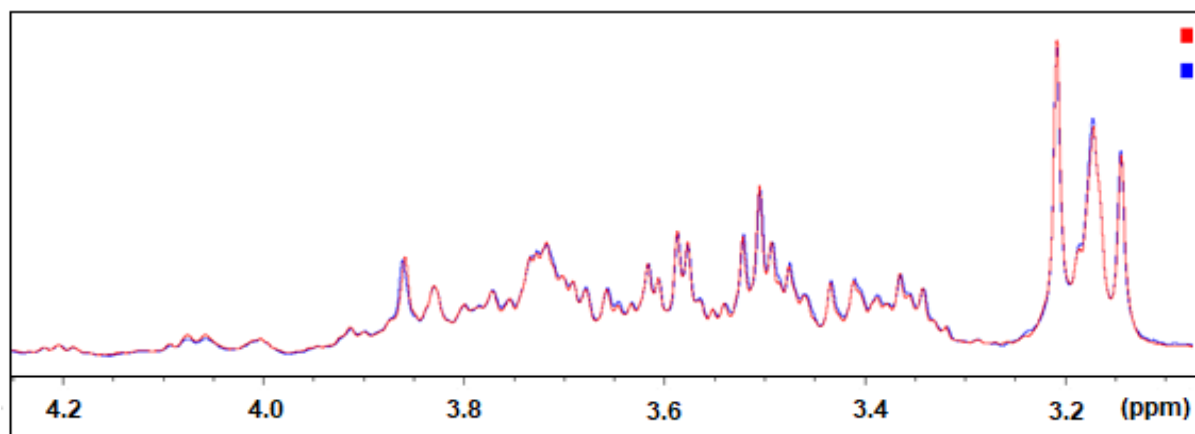
Figure 76 *No significant changes in the spectra even after two hours of experiments are reported. Here, the superposition of two spectra: red, at time 0 and the blue, after 2 hours.*

Conventionally, HR-MAS metabolomic studies are carried out by using CPMG filtered spectra at a relatively fast-spinning rate (4 kHz). The rotor synchronization does not seem to be a drawback and leads to relatively short inter-pulse delays (250 µs) that can be considered as

sufficient to refocus the *J*-modulation, even if this refocusing is imperfect. At low spinning rates (1 kHz, for example), rotor synchronization leads to a long inter-pulse delay (1 ms) and the PROJECT sequence is used to generate spectra without "*J*-modulation artifacts". $T_{1\rho}$ filtered experiment can be a simple and efficient alternative since there is no precession during the evolution period of the magnetization and therefore no *J*-modulation artifacts. However, some properties related to the dependence of $T_{1\rho}$ to the nutation (ν_1) and the rotation (ν_r) frequencies, in addition to the obvious contribution of slow dynamic processes, have to be taken into account. This dependence has been widely noted on several applications are using $T_{1\rho}$ for the determination of the local dynamics of proteins by MAS.[174–176]

Recently, Krushelnitsky *et al.*[177] published a comprehensive article that presents the MAS dependence of the relaxation rate $R_{1\rho}$, which was modulated by correlation times of slow motions, from 10 μs to 1 ms, green, red, and black curves in **Figure 77** which depicts the $R_{1\rho}$ evolution as a function of spinning rates and correlation times.
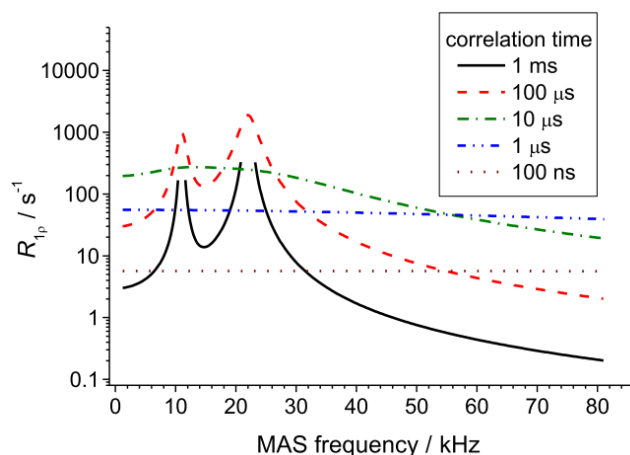
Figure 77 *Analytical calculation of the on-resonance ^{15}N $R_{1\rho}$ MAS dependence at different correlation times of motion of a $^{15}N–^1H$ spin pair. Reprinted with permission from A. Krushelnitsky et al.[105] Copyright 2014 Elsevier.*

In semi-solid samples, one can expect a contribution of both isotropic and anisotropic interaction to the relaxation mechanisms.[178] To evaluate this behaviour, we measured the $R_{1\rho}$ at the two spinning rates, 1000 and 4000 Hz, commonly used in HR-MAS metabolomics, for selected B$_1$ amplitudes. We report in **Figure 78** the relaxation rates $R_{1\rho}$ ($= 1/T_{1\rho}$) of some

selected metabolites for one randomly chosen sample (control-3) to assess its evolution as a function of B_1 expressed by the nutation frequency ($\upsilon_1 = \gamma B_1/2\pi$) for the two MAS rates, 1000 Hz (**Figure 78.a**) and 4000 Hz (**Figure 78.b**).
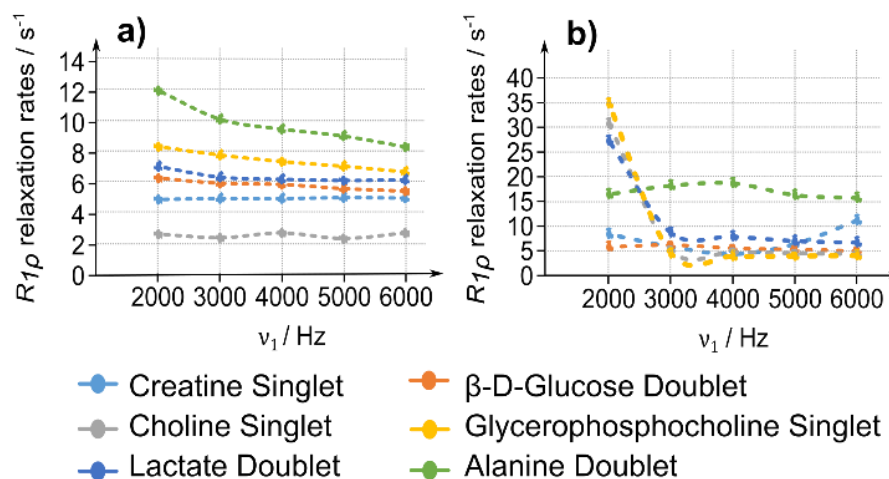
Figure 78 $R_{1\rho}$ *as a function of* ν_1 *for liver control sample (C3) relative to six different peaks of six metabolites (a) at* ν_r *= 1000 Hz and (b) at* ν_r *= 4000 Hz. Dotted lines are used to help visually track the progress of* $R_{1\rho}$.

The general observation concerns the MAS dependence of the relaxation rates, which confirms the behaviour reported in the literature in the solid state.[177] If we focus on $\nu_1 = 4000$ Hz, which was defined as optimal in the liquid study (as a better compromise for a moderated amplitude, sufficient spectral coverage, and better resolution), one can analyse quantitatively these evolutions (**Table 34** and **Table 35**).

Table 34 $R_{1\rho}$ (s^{-1}) *evolution of some chosen metabolites, from liver sample C3, randomly chosen, as a function of spinning (* ν_R *) and nutation (* ν_1 *) frequency.*

ν_1 (Hz) / Metabolites	ν_R = 1000 Hz					ν_R = 4000 Hz				
	6000	5000	4000	3000	2000	6000	5000	4000	3000	2000
β-D-Glucose Doublet	4,92	4,95	4,88	4,89	4,83	11,84	7,07	5,03	6,49	9,05
Creatine Singlet	5,38	5,54	5,81	5,89	6,22	5,66	5,95	6,19	6,85	6,45
Glycerophosphocholine Singlet	2,64	2,29	2,64	2,33	2,56	5,06	5,19	5,47	5,85	31,32
Choline Singlet	6,63	6,99	7,302	7,72	8,28	4,70	4,63	4,59	5,12	35,39
Alanine Doublet	6,11	6,07	6,15	6,28	6,98	7,35	7,66	8,56	9,32	27,91
Lactate Doublet	8,27	8,97	9,42	10,07	11,95	16,35	16,92	19,31	18,83	17,09

Table 35 $R_{1\rho}$ *evolution of some chosen metabolites, from liver sample C3, as a function of spinning rate (v_R) for v_1 = 4 kHz.*

	$v_R = 1000$ Hz	$v_R = 4000$ Hz	$R_{1\rho}$ evolution
β-D-Glucose Doublet	4,88	5,03	↑
Creatine Singlet	5,81	6,19	↑
Glycerophosphocholine Singlet	2,64	5,47	↑
Choline Singlet	7,302	4,59	↓
Alanine Doublet	6,15	8,56	↑
Lactate Doublet	9,42	19,31	↑

For certain metabolites, probably depending on the slow dynamic contribution, $R_{1\rho}$ increases with the MAS rate (*e.g.*, lactate doublet and glycerophosphocholine singlet), for others, it remains substantially the same (creatine singlet and β-D-glucose doublet), and for some others, it decreases (*i.e.*, choline singlet). These expected results, due to the heterogeneity of the tissues analysed in HR-MAS, do not constitute a drawback for metabolomic applications. Indeed, the behaviour of each metabolite is reproducible through different samples. However, the stability of the relaxation values around the chosen v_1 (4000 Hz) is critical. For both MAS rates, we found that around this value, the relaxation rates would not be drastically affected by the expected slight variation in B_1 due to eventual different probe tuning from one sample to another. To assess qualitatively the reproducibility, we recorded the T_2- and $T_{1\rho}$-edited spectra for three samples taken from the same control liver (**Figure 79**).
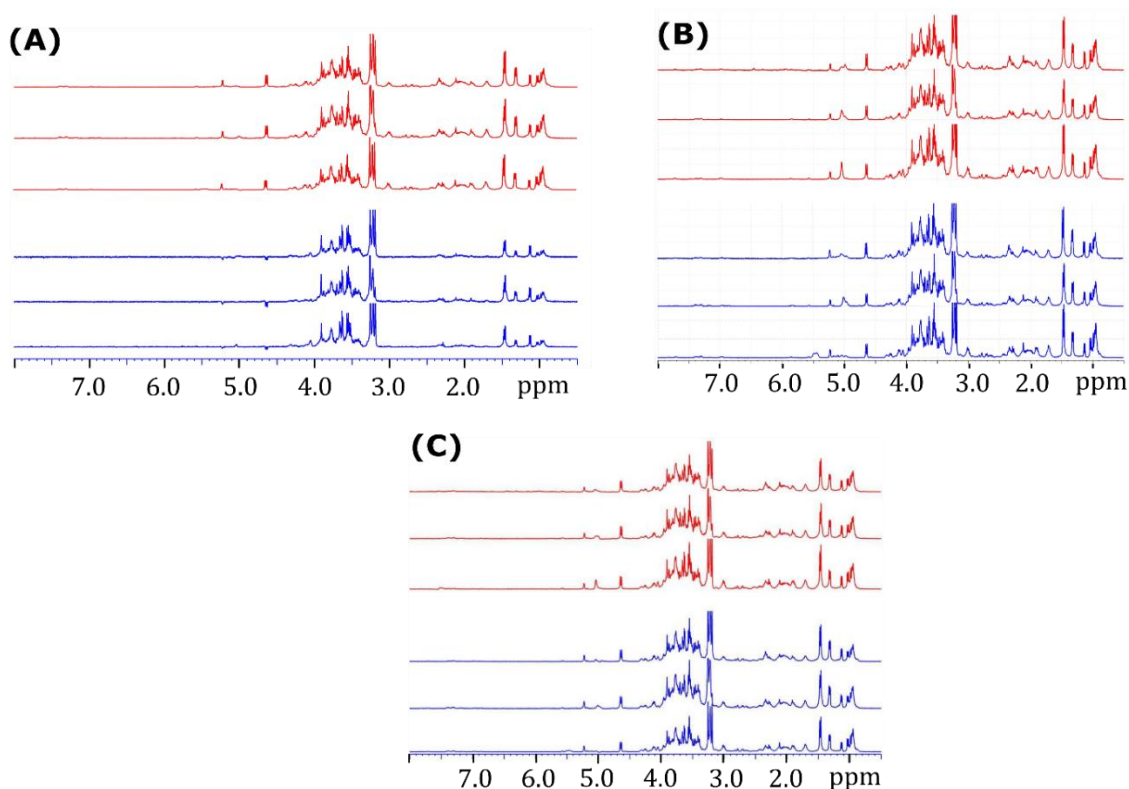
Figure 79 - T_2 and $T_{1\rho}$ edited spectra for three samples taken from the same control liver, sample C2. (**a**) CPMG, (**b**) $T_{1\rho}$, (**c**) PROJECT. In red the spectra acquired at v_R = 4000 Hz, in blue at v_R = 1000 Hz.

The $T_{1\rho}$ and PROJECT filtered spectra are identical whatever be the MAS rate (v_r = 1000 and 4000 Hz), while the CPMG-filtered spectra present *J*-artefacts at v_r = 1000 Hz, which result in negative signals and an intensity loss, especially between 1.5 and 3 ppm (**Figure 79.a**).

III.1.11 Statistical Analysis on Liver Samples

The spectra obtained with each pulse sequence (CPMG, PROJECT and $T_{1\rho}$) have the same filtration effect but the profiles show predictable intensity differences due to the different mechanisms involved in each relaxation filter.[179] The impact of these differences on the statistical processing obviously depends on the reproducibility of the artefacts of each sequence Another important experimental consideration for the $T_{1\rho}$ is to avoid the synchronization of the

spinning frequency ν_r with the spin-lock frequency ν_1. Indeed, this leads to a recoupling of the Chemical Shift Anisotropy (CSA), dipolar interactions, and magnetic susceptibility effects, which results in a loss of resolution.[180,181] To avoid these contributions, we used a spin-lock amplitude of 4.3 kHz for both used spinning rates, 4 kHz and 1 kHz. **Figure 80.a** presents an OPLS-DA plot of all the spectra obtained with a MAS = 4 kHz. It shows that the CPMG and the PROJECT spectra are relatively similar and significantly different from the $T_{1\rho}$ filtered one. Interestingly, at low spinning rate, 1 kHz (**Figure 80.b**), the synchronization of the inter-pulse delays (1 ms) lead to significant discrimination between CPMG and PROJECT filtered spectra, due to the *J* modulation effect, which is not compensated with the CPMG sequence.
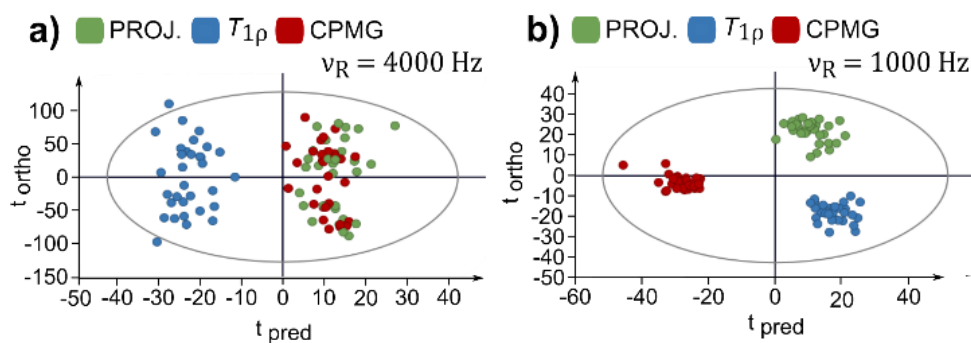
Figure 80 OPLS-DA score plots performed on all liver samples including the two groups C and HFS obtained for each sequence (CPMG vs. PROJECT vs. $T_{1\rho}$) at ν_r =4 kHz (**a**) and ν_r =1 kHz (**b**).

These differences can be assessed in a relevant way, directly on the spectra, but, in addition, statistical tools allow to evaluate the overall "inter-sequence" differences and the "intra-sequence" reproducibility.

The OPLS-DA multivariate analyses on the two groups, control against HFS, are depicted in **Figure 81** and **Figure 82.** They enable us to evaluate the separation capability and the steadiness of $T_{1\rho}$ and T_2 filtered spectra obtained at the two spinning rates, 1 kHz and 4 kHz.
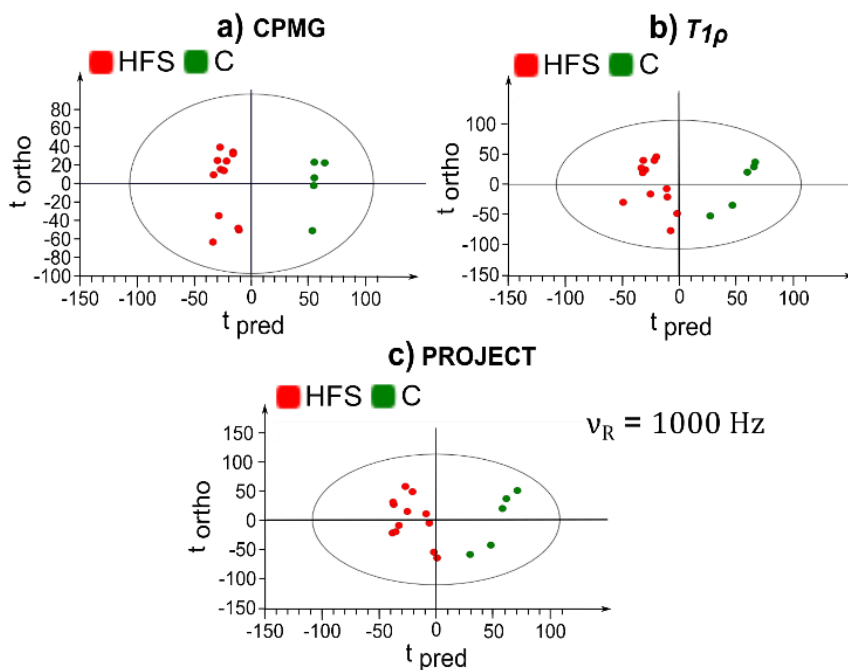
Figure 81 *Multivariate analysis using OPLS-DA performed on liver samples for the two groups: C and HFS obtained for each sequence: CPMG (a), $T_{1\rho}$ (b) and PROJECT (c), at vr =1 kHz.*
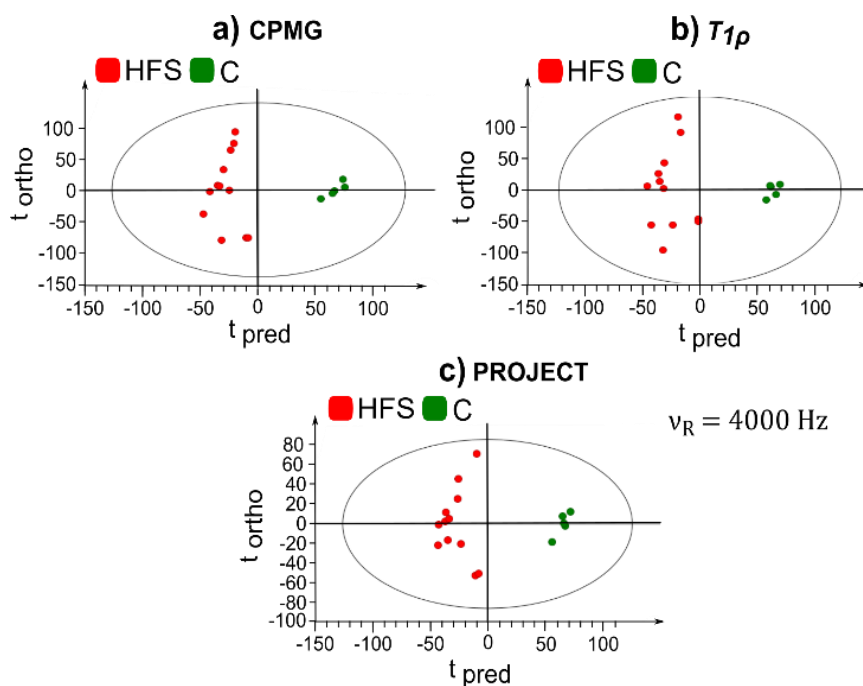
Figure 82 Multivariate analyses using OPLS-DA performed on liver samples for the two groups: C and HFS obtained for each sequence: CPMG (a), $T_{1\rho}$ (b) and PROJECT (c), at vr =4 kHz.

Whatever the spinning rate, the main observation is that the discrimination between the two groups is clear and the score plots obtained with the $T_{1\rho}$ filter provide similar statistical parameters to those ones obtained with PROJECT and CPMG sequences (**Table 36**).

Table 36 *The OPLS-DA $R^2{}_Y$, $Q^2{}_Y$ and p-values for HFS vs. C samples of livers without one outlier (with N= number of the samples and A= number of components to fit the model) using Unit Variance (UV) scaling.*

	N	A	$R^2{}_Y$	$Q^2{}_Y$	*p*-value
ν_r =1000 Hz					
CPMG	17	1+1+0	0,966	0,770	0,00083
$T_{1\rho}$	17	1+1+0	0,869	0,673	0,0062
PROJECT	17	1+1+0	0,857	0,650	0,0090
ν_r = 4000 Hz					
CPMG	17	1+1+0	0,947	0,798	0,00039
$T_{1\rho}$	17	1+1+0	0,923	0,793	0,00045
PROJECT	17	1+1+0	0,943	0,797	0,00041

The observed intragroup variation for the HFS group in each experiment can be explained by taking into account the inhomogeneity of the response of complex organisms to an external imbalance, here a food difference.[182,183] This effect is observed on several parameters: overall weight of animals, gain weight, normalized liver weights, Homeostasic Model Assessment of insulin resistance (HOMA-IR) levels, serum glucose, triglycerides, cholesterol and insulin levels. In the previous article, which dealt with these samples,[167] we clearly noted an increase of the standard deviation for all these parameters when we pass from controls to an increasing introduction of the number of external stimuli or imbalance (standard diet, HFS diet, HFS diet plus medication) (**Table 37**).

Table 37 Results of serum biochemical analysis and body weight after 10 weeks of diet. Table duplicate from the paper by Tranchida et al.[166] Data of an additional group (HFS + C) are added to illustrate the increase in standard deviation by the introduction of an external factors: HFS dietary imbalance, then HFS + C supplementation (or medication). The heterogeneity of the response to each stimulus results in an increase in the standard deviation.

HFS high fructose and saturated fatty acids, C curcuma (administration of hydroalcoholic extract of tumeric 100 mg/kg/day). Relative liver weight is defined as liver weight divided by body weight. Values are mean ± S.E.M (n = 6-12 rats/group).

Group	Controls	HFS	HFS+C
Body Weight (g)	435.17 ± 20.74	438.42 ± 31.85	444.75 ± 50.78
Gain Weight (g)	245.16 ± 8.47	248.42 ± 9.19	254.75 ± 14.66
Relative liver weight	0.0262 ± 0.00078	0.0310 ± 0.0010 **	0.0334 ± 0.0016 **
HOMA-IR	1 ± 0.33	10.79 ± 1.89 **	6.97 ± 2.05 *
Insulin (μg/l)	0.48 ± 0.13	3.34 ± 0.50 *	2.22 ± 0.69 *
Glucose serum (g/l)	1.07 ± 0.185	1.82 ± 0.41 *	1.80 ± 0.44 *
Total cholesterol serum (g/l)	0.60 ± 0.02	0.68 ± 0.04 *	0.66 ± 0.06
Triglycerides serum (g/l)	0.435 ± 0.15	0.878 ± 0.22 *	1.114 ± 0.36 *

*$P < 0.05$ vs. the control
**$P < 0.01$ vs. the control.

HOMA-IR: Homeostasic Model Assessment of Insulin Resistance.

This dispersion is usually lower or even absent for the control groups.[167] However, whatever the sequence used, this intra-group reproducibility is not found at ν_r =1 kHz, as we can see in **Figure 81**, because of the non-reproducible sidebands present in the spectra, depending on the sample preparation. At ν_r = 4 kHz, the absence of the sidebands allows to retrieve the intra-group reproducibility for the control group (**Figure 82**).

These results show that the artefacts inherent in each sequence are reproducibly replicated over the samples, and the multivariate analysis allows to detect the discrimination between the two groups, including the CPMG sequence at ν_r = 1000Hz, which presents J-artefacts (negative signals and an intensity loss especially between 1.5-3 ppm) due to the long inter-pulse delay (1 ms) used for rotor synchronization (**Figure 81**). However, CPMG sequence must obviously be excluded at ν_r = 1000 Hz, since the quality of the spectra does not allow an accurate

determination of biomarkers. The $T_{1\rho}$ and PROJECT filtered spectra are identical for the two MAS spinning rates (v_r = 1000 Hz and 4000 Hz) and the loading plots provide the same profiles with comparable correlation coefficients (**Figure 84** and **Figure 83**). However, the loading plot for the CPMG at v_r = 1000Hz (**Figure 83.a)** is obviously different and is unsuitable for the extraction of discriminating metabolites.
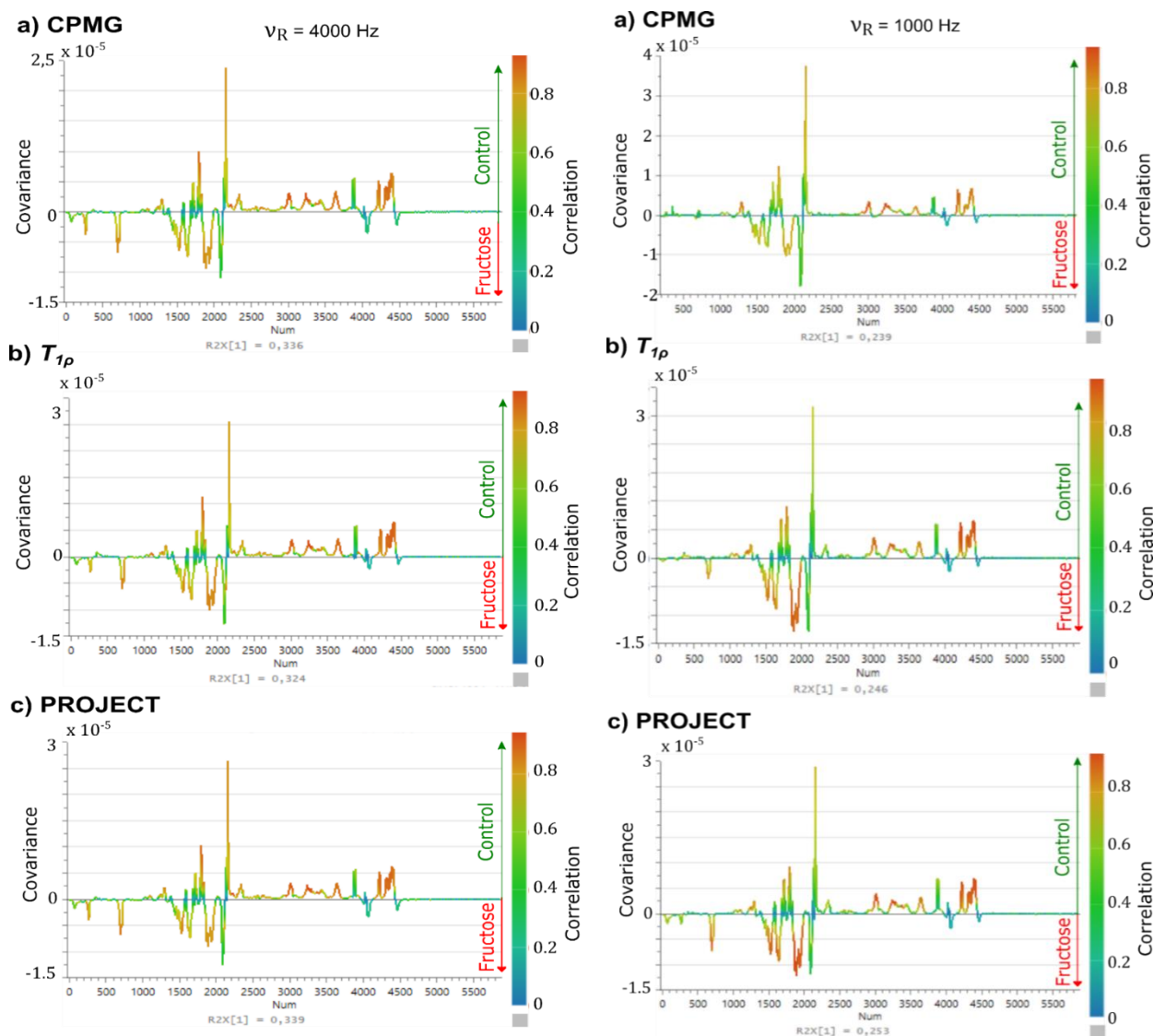
Figure 84 loading plots of the OPLS-DA models of Figure 82 for a) CPMG, b) $T_{1\rho}$, c) PROJECT at v_R = 4000 Hz for liver samples of control and HFS (fructose) groups.

Figure 83 Loading plots of the OPLS-DA models of Figure 81 for a) CPMG, b) $T_{1\rho}$, c) PROJECT experiments at v_R = 1000 Hz for liver samples of control and HFS (rich in fructose diet) groups.

III.1.12 Conclusions for $T_{1\rho}$ Assessment

In this section, we have shown the application of the $T_{1\rho}$ filter in metabolomics on liquid samples (serum) and on tissues (liver). Its implementation is simple and does not require calculations for the adjustment of the duration of the filters, nor synchronization whatever the MAS spinning rates. Despite its advantages, its use remains "rejected", probably due to the overheating effect which would be caused by the spin-locking. For the filter duration (hundreds of ms) and also the amplitudes used in metabolomics, up to 5 kHz, the heating is not a drawback. It is important to note that this work relates to spectrometers operating at 400 MHz and 600 MHz. If metabolomic experiments with a $T_{1\rho}$ filter were to be performed at higher magnetic fields, the power of the spin-lock may not be sufficient to cover the entire spectrum. To avoid the offset effects, it will be necessary to use higher amplitudes, that, taking into account the metabolomic duration of the filters, should not have a deleterious effect. Moreover, we showed that if moderated offset effect is present, the analysis of the results leads to the same observations: the inter-group discrimination, and the extraction of the discriminant spectral regions from the loading plots. However, the developments and the implementation of the offset methods used in solid-state NMR[184] to enhance the total effective field of the spin-lock represent an interesting prospect for wider spectral covers in metabolomics.[185]

III.2. Application of the 2D Maximum Quantum NMR for the characterization of enzymatic reaction medium

In this section, we present a simple analytical method for rapid characterization of reaction mixtures involving enzyme complexes, using Maximum Quantum (*MaxQ*) NMR, accelerated with the Non-Uniform Sampling (NUS) acquisition procedure. Specifically, this approach enables, in a first analytical step, to count the components present in the samples. We show, using two different enzymatic systems that, in this case, the implementations imply precautions related to the short relaxation times, in particular, for metalloenzymes or paramagnetic catalysts. Under these conditions, the 2D MaxQ experiment can be used qualitatively, for the assignment of the compounds. Finally, the combination of MaxQ NMR and diffusion, which leads to a 3D chart, greatly improves the resolution and offers an extreme simplification of the spectra while giving valuable indications on the affinity of the enzymes to the different compounds present in the reaction mixture.

III.2.1 Quick Review of Enzyme Activity Monitoring by NMR

NMR has been widely used to monitor enzymatic activity by recording the evolution of the spectra of substrates and/or products to characterize enzymatic processes that are complex, dynamic, and time-dependent. For complex systems involving the coexistence of multiple compounds (substrate, final product, and various intermediates), the identification and quantification can be a more arduous task. Thanks to the quantitative linear response, one dimensional ^{1}H and hetero-nuclei (^{13}C, ^{31}P, ^{19}F...)[52][186] NMR have been the most used to record the delicate and fast variations in concentrations of substrates and products as a function of time.[187][188] The methodological developments associated with recent studies have undergone a

significant renewal and provided versatile protocols, as the real-time NMR experiments,[189] [190] [191] to monitor in a fast and reproducible manner enzyme kinetics[192] and to estimate, under favourable conditions, the Michaelis constant.[193] [188] In general, the effective use of quantitative ^{1}H NMR depends on the speciation of one or several well-defined peaks from crowded spectra for each component in the reaction mixture, even if the information has to be extracted from overlapping peaks. In this context, the contribution of NMR methods designed for mixture analyses is valuable and obvious. The algorithms for mathematical deconvolution of superimposed peaks were successfully transposed to the real-time NMR to identify the metabolites and propose an enzymatic reaction network.[193] The most common NMR techniques used for mixture analysis are TOtal Correlation SpectroscopY (TOCSY) and Diffusion Ordered SpectroscopY (DOSY) (that brings the closest similarity to chromatography). As it was discussed in paragraph I.1.2.4., in this experiment, molecules are labeled by their molecular diffusion, which is roughly determined by the molecular size.[46] Thus, the spectra of the components can be extracted if the differences in the molecular diffusions are large enough.[194] [52] [53] [54] [48] [47] Otherwise the resolving power of DOSY will be restricted.[195] The recent developments of the three-dimensional NMR sequences aimed to increase the resolving power of DOSY through a third dimension to resolve the overlapping resonances.[196] [197] [198] [199] [200] but the wide use of these methods has been thwarted due to the long acquisition time required for recording several dimensions.

III.2.2 Maximum-Quantum (MaxQ) NMR for the Analysis of Complex Mixtures

^{1}H Maximum-Quantum (MaxQ) NMR is a recently introduced approach for mixture analysis. It provides high-resolution spectra, in favorable cases up to tens of molecules.[201] MaxQ-NMR is a 2D experiment that aims at isolating, in a complex NMR spectrum, the signals of molecular fragments, carrying at least a specific number of protons, *p*. In the indirect

dimension, all of these signals contribute to a single Maximum-Quantum resonance, as in an ideal well-resolved 2D DOSY spectrum. MaxQ order corresponds, in fact, to the higher quantum transition leading to a singlet in the second indirect dimension. Moreover, these signals are dispersed in the horizontal direct direction according to their chemical shift and scalar coupling corresponding to the single quantum order ($p=1$). This technique can filter out the signals of several molecules, and thus simplify the spectra. Beyond the 2D-MaxQ NMR, the 3D-MaxQ-DOSY NMR leads to simplified spectra.[202] [203] MaxQ-NMR applications until now have focused on food chemistry (polyphenols) and environmentally relevant species (polyaromatic molecules).[204] [62] [205] [206] In this study, we go beyond exploring the application of 2D-MaxQ NMR and 3D-MaxQ-DOSY NMR to simplify the spectra of the enzymatic reaction mixtures. Two enzymatic systems were used as case studies to assess this approach. The first system involved a fungal Copper Radical Alcohol Oxidase assisted by two accessory enzymes with 4-pyridinemethanol as substrate. The second system relies on a photo-induced system using a catalyst.

III.2.3 Enzymatic systems under investigation

In the first system the enzyme, a fungal alcohol oxidase (Copper Radical Alcohol Oxidase, *Cgr*AlcOx), carries out the oxidation of 4-pyridinemethanol (the substrate), to an aldehyde and a carboxylic acid. The major expected products/intermediate are given in **Table 38**.

Table 38 *Probable compounds of the first enzymatic system. CrAlOx enzyme from PDB database (on the right) which catalyses the oxidation of the alcohol to the corresponding aldehyde.*

4-pyridinemethanol   Diol   Aldehyde

CrAlOx
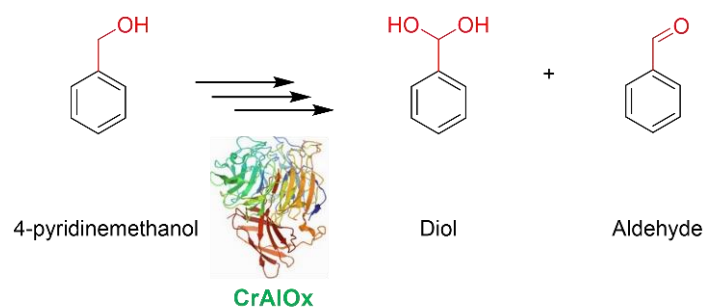
The enzyme employed in the second system is a laccase, a multi-copper oxidase, presents in plants, fungi, and bacteria, which oxidizes the substrate through the active site of four copper centres. The substrate used here is the para-styrene sulfonate (pSS), which is transformed in epoxide giving an opened diol species later, and then becomes an aldehyde. The major expected products/intermediates are given in **Table 39**.

Table 39 *Compounds of the second enzymatic system with the substrate para-styrene sulfonate, the intermediates and the final expected product.*

pSS   Epoxide   Diol   Aldehyde

LAC3
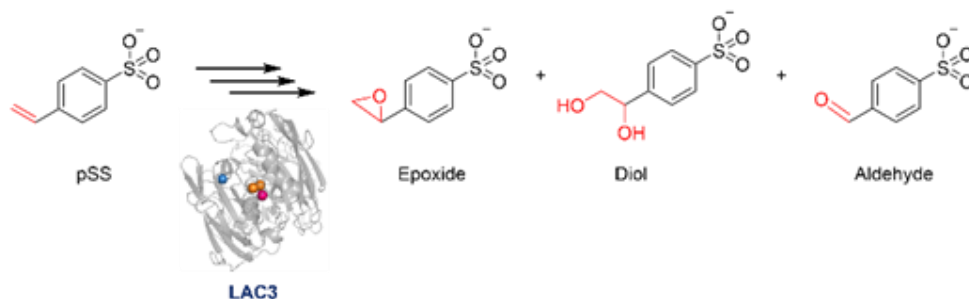
III.2.4 NMR Analysis of the Enzymatic Reaction Medium

The 1D ^{1}H spectra of both enzymatic mixtures present several signals in the aromatic regions, expected from the different compounds (intermediates) present in addition to the initial substrate and the final product. **Figure 85** shows the 1D proton spectrum of the first reaction mixture after 16 hours of reaction in presence of 4-pyridinemethanol as substrate.
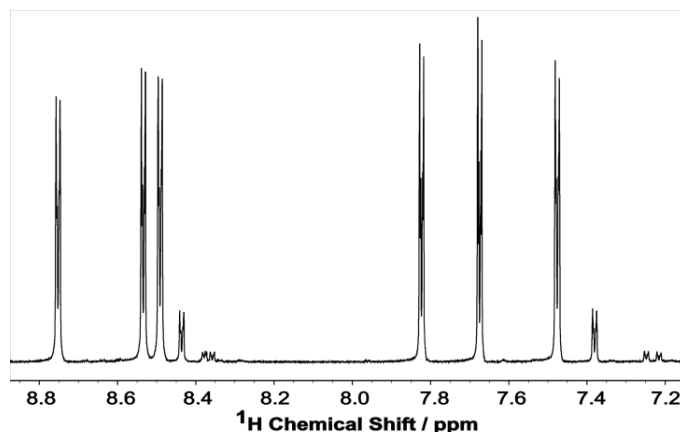
Figure 85 *1D ^{1}H spectrum of the first enzyme system with the 4-pyridinemethanol as substrate, with a zoom in the region between 8.8 and 7.2 ppm.*

As explained before, the ^{1}H Maximum-Quantum NMR is a 2D experiment that isolates the signals of the same molecular fragments. Depending on the number of protons in the coupled spin system, the excitation of the highest coherence order (MaxQ) provides a singlet in the indirect dimension correlating all the peaks involved in this spin system. In this study, the substrates, 4-pyridinemethanol and para-styrene sulfonate (pSS), as well as the products and intermediates (illustrated in **Table 38** and **Table 39**) are di-substituted aromatic molecules, with a maximum excitable coherence of 4 quanta *(p = 4)*. Therefore, the 4Q multiple quantum spectra will allow to isolate the peaks belonging to the different molecules in the mixture.

These MaxQ spectra (2D MQ/SQ correlation) were acquired with a standard sequence 90_x-$\tau/2$-180_x-$\tau/2$-90_ϕ described in detail in previous works,[62,201,207] the coherence order selection was carried out by two pulse field gradients around the last 90° pulse, by selecting the intensity of the second PFG equal to *p* times the first one.[208] The interval d_2 was optimized to obtain the highest signal intensity, for the 4Q-1Q spectra, this corresponded to 90 ms and 55 ms for the first and second samples, respectively (**Figure 86**).
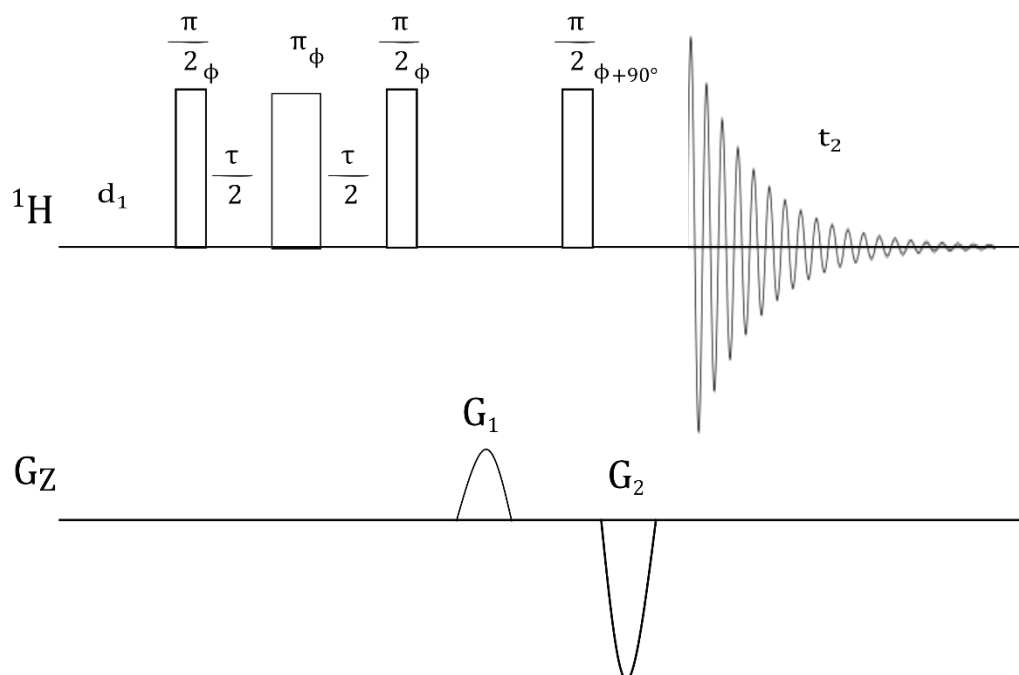
Figure 86 *2D MaxQ pulse sequence. The duration of the preparation period τ was optimized to obtain as uniform as possible MQ excitation of the desired coherence order. The ratio of the coherence selection gradient pulses was chosen to fulfil G2 = p × G1, where p is the MQ coherence of choice.*

In the 2D MQ NMR, the multiple quantum coherences are excited using the pulse sequence in **Figure 86**. After the last ($\pi/2$) pulse, the amplitude of the antiphase terms produced by the coupling Hamiltonian is determined by the intensity of the selected coherence orders which are modulated by (**Equation 3.2**):[209]

$$2^p \prod_{i=2}^{p-1} \sin(\pi J_{IX}\tau) \; exp^{-\left(\frac{\tau}{T_{2i}^*}\right)} \qquad (3.2)$$

The amplitude of the different coherence orders is strongly modulated by the main delay of the preparation period, τ.[187] In a spin system with different coupling constant J, the delay τ, which provides the maximum intensity, corresponds to the reciprocal of the smaller coupling constant (J_{small}), $1/2J_{small}$. Theoretically, in an aromatic spin system, this corresponds to J about 1 Hz, hence $\frac{\tau}{2} = 0.5 \; s$. It has been shown that for free small molecules the sinusoidal dynamics along the τ delay is hampered by the exponential decrease due to the relaxation[188] and that $\frac{\tau}{2}$ around 0.3 s is most often found as the optimal value.[210] However, this feature has never been assessed

in the presence of large entities where the relaxation times decrease drastically. The evolution of the intensities of the 4Q signals as a function of the $\frac{\tau}{2}$ delay for the two pure substrates and the reaction mixtures are depicted in **Figure 87**. As expected, the evolutions in the presence of the enzymatic complexes are different from those ones of the free molecules. According to **Equation 3.2** the sine evolution of the 4-quanta coherence intensity will be thwarted by the relaxation decay leading to signal attenuation. The ideal and uniform excitation is rarely achieved because of this effect and the good compromise is to get as close as possible to the optimal value to access distant couplings without signal loss. For the free molecules (**Figure 87.c** and **Figure 87.d)** the second vertex of the sinusoid provides a relatively uniform excitation with the $\frac{\tau}{2}$ delay of 230 ms and 180 ms for the first and the second substrate, respectively. For the reaction mixtures (**Figure 87.a** and **Figure 87.b**) the relaxation decay is more manifest. However, for the first system (**Figure 87.a)**, the second vertex remains visible at $\frac{\tau}{2} = 210 \, ms$ unlike the second system (**Figure 87. b**) for which the only exploitable vertex is the first at $\frac{\tau}{2} = 55 \, ms$.
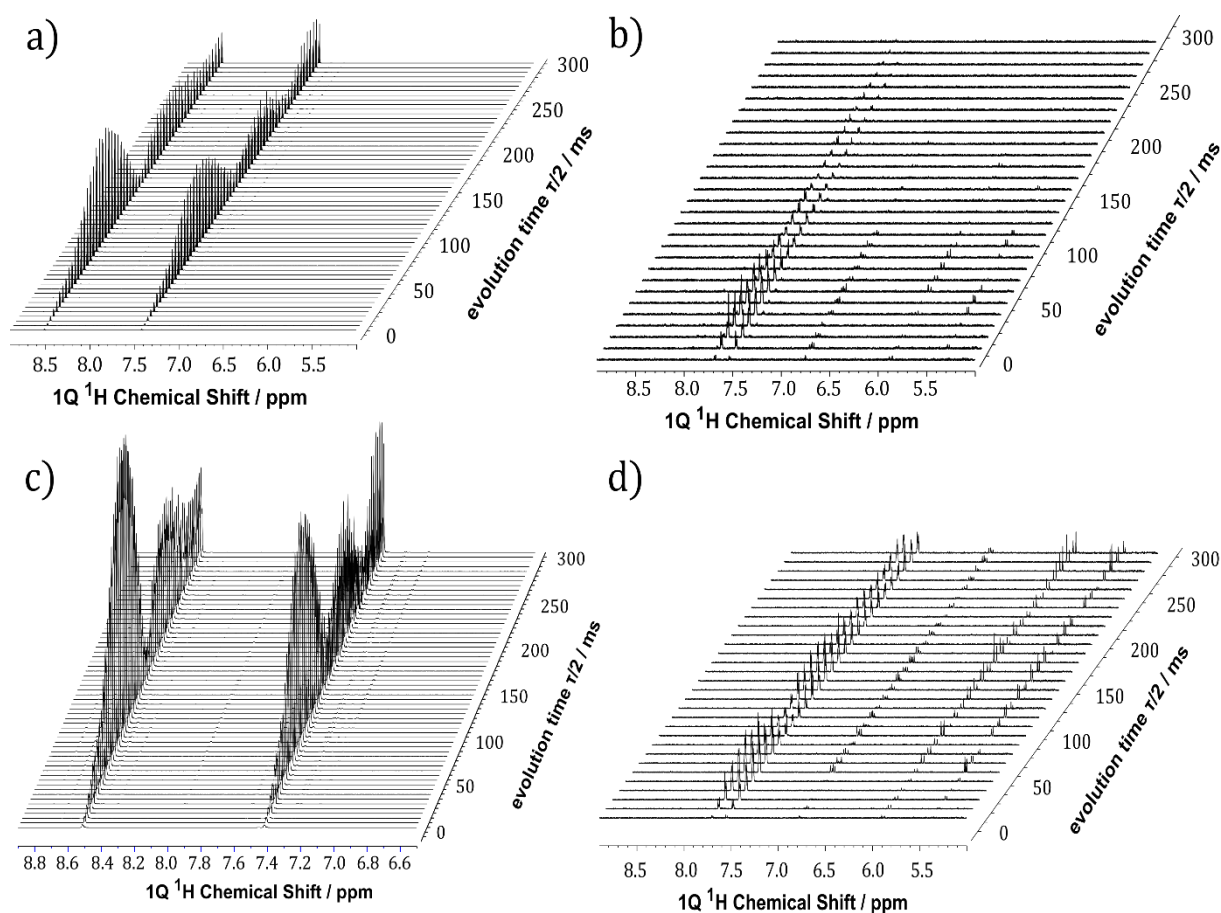
Figure 87 *Evolution of the 4Q-filtered signals, displayed in absolute mode, of the reaction mixtures. **a)** first system with the enzyme, **c)** first system only with the substrate, **b)** second system with the enzyme, **d)** second system only with the substrate.*

The interactions between the compounds and the large protein complexes, whether specific or not, lead to a drastic reduction of the relaxation times. However, depending on the sample used, the behaviours are different. We retrieve the sinusoidal oscillation for the first system **(Figure 87.a)** but it is absent for the second one where the relaxation contribution seems very important leading to rapid extinction of the signal, probably due to the paramagnetic contribution from the multi-copper oxidase enzyme, the laccase, and the ruthenium photo-catalyser (**Figure 87.b**). These observations show that a standard implementation of MQ NMR cannot be envisaged to characterize the enzymatic reaction mixtures. The choice of the optimal preparation time, τ, should be guided visually for each sample.

Figure 88 and **Figure 89** show the 2D 4Q-1Q spectra for the two reaction mixtures obtained with the sequence **Figure 86**. These spectra provide a fingerprint in which the visual

enumeration of molecules is simplified since the peaks belonging to the same molecular fragments are aligned on the same frequency in the indirect dimension, and we can count six and four different components for the first (**Figure 88**) and the second reaction mixture (**Figure 89**), respectively.
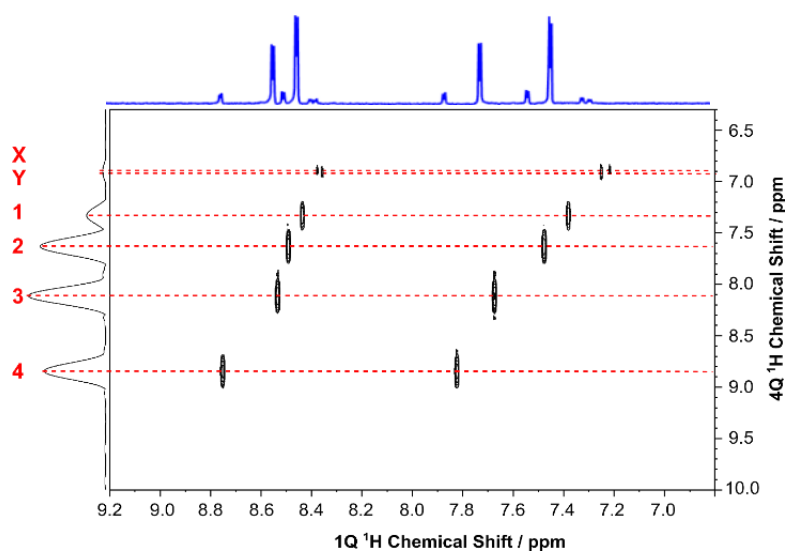
Figure 88 *2D 4Q-1Q spectrum of the first system with a NUS at 25 % acquired with TD1=128, NS=32 with a total experimental time of 30 minutes.*

An interesting feature can be highlighted. When certain products or intermediates are present in low concentration, they can be confused with ^{13}C satellites in the 1D ^1H spectra. This ambiguity is resolved in the 2D MaxQ spectra where only the peaks of the compounds are recorded.
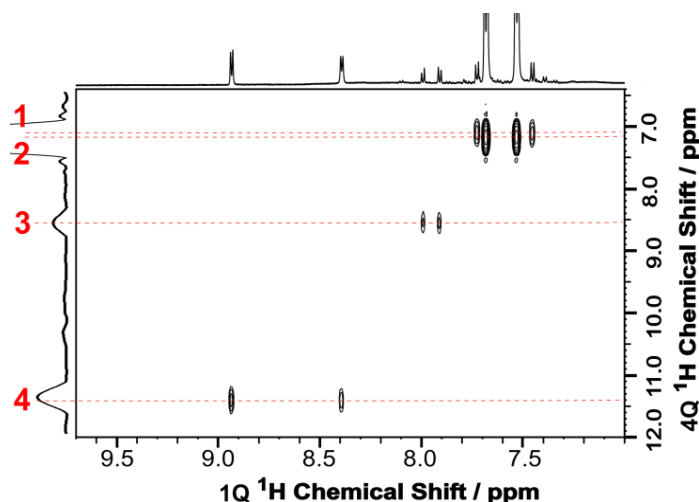
Figure 89 *2D 4Q-1Q spectrum of the second system with a NUS at 25 % (compression factor=4) acquired with TD1=128, NS=32 with an experimental time of 30 minutes.*

III.2.5 Acceleration and Resolution Improvement

For dynamic systems, it is important to propose and perform experiments with short acquisition times. To reduce the experiment time without information loss, we may apply the standard NUS acquisition procedure available within the Topspin software. It is known that the assessments of the correct NUS spectra reconstruction are qualitative and closely linked to the signal-to-noise ratios.[211] We did not resort to this visual evaluation, since this work has been widely carried out in the literature, and it is now admitted that a compression ratio of 4 (NUS at 25%) allows producing faithful spectra in a reasonable experimental time without significant loss of resolution or signals while avoiding artifacts. For each sample, the experiment time of the two-dimensional 4Q-1Q spectrum was reduced to 34 min. Another application of the NUS acquisition procedure is to increase the resolution in the indirect dimension without experiment time cost by increasing the number of points in the indirect dimension with a factor equal to the compression rate. **Figure 90** shows the NUS 2D 4Q-1Q spectrum of the laccase reaction mixture (second system) recorded with 256 points in F_1 and a compression rate of 4. An increased resolution is obtained and enables the two peaks around 6 ppm, in the indirect

dimension, to be well resolved. However, some compression artifacts appear (red circle in **Figure 90**).
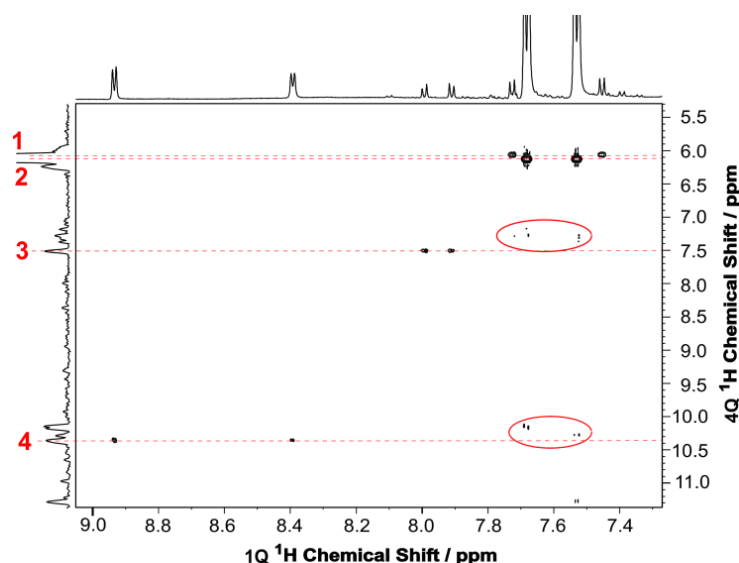
Figure 90 *2D 4Q-1Q NUS (25%) of the second system using 256 points (TD1) in the indirect dimension, obtained in 48 minutes. Standard experiment time is 3.25 hours. Compression artefacts are showed inside the red circles.*

III.2.6 2D DOSY and 3D MaxQ-DOSY

In this framework, the effectiveness and the power of DOSY NMR for direct identification of mixture composition are well established.[48,54] Provided that both mobility and spectral resolution are sufficient, it allows the extraction of NMR spectra of pure compounds. On the other hand, the resolving power of DOSY has been limited by technical and signal processing difficulties. Indeed, much of the research in the development of better DOSY experiments focuses on this latter aspect. When a numbering purpose is sought during a first analysis step, DOSY NMR can also be an appropriate method. This feature is explored in this section. The DOSY spectrum of the *Cgr*AlcOx reaction mixture (first system) obtained using the standard convection compensated pulse sequence is presented in **Figure 91.** The first interesting finding that can be highlighted concerns the non-expected resolution due to the mobility which is sufficient to allow the detection of the six compounds and extract their

diffusion coefficients. Given the identical molecular size of these compounds which have only undergone chemical modifications at the main functions (alcohol - aldehyde- acid-...), the diffusion coefficients should be very close and would not allow isolating the compounds. The only entity presents in the mixture which can affect the diffusion coefficients in a different way is the protein complex, which can differentially interact with the compounds according to the affinity.
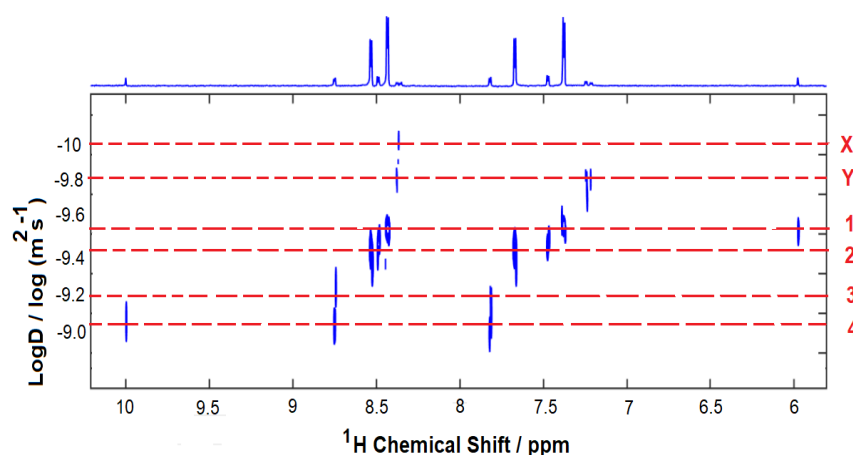
Figure 91 *2D DOSY spectrum for the first system using a standard convection compensated pulse sequence (Bruker notation: dstebpgp3s). Processed with DOSY Toolbox (GNU General Public License) with 512 points.*

Standard DOSY spectra were recorded by a conventional convection compensated pulse sequence, based on the stimulated echo and incorporated bipolar gradient pulses, and an Eddy current delay (BPP-LED).[199] The shape of all gradient pulses was Smoothed Square and the LED delay was 5 ms. The diffusion delay Δ was set at 100 ms, the gradient strength, g, was linearly incremented in 32 steps from 2% to 98% of its maximum value with a duration of 1 ms, and 8 scans were recorded. Data were processed using Dosy Toolbox (GNU General Public License)[212] with 256 points for the inverse Laplace diffusion dimension.

Several three-dimensional NMR sequences have been proposed to increase the spectral resolution of DOSY through a third dimension to further spread out overlapping

resonances.[199,197,213,196] Recently, we developed 3D MaxQ-DOSY that associates the two "*demixing*" power of each method (MaxQ and DOSY) and offers a more simplified layout.[63] The diffusion measurement using multiple-quantum coherences was demonstrated in the cases of scalar, dipolar, and quadrupolar coupling.[214–218] It has been shown that if the dephasing information resulting from translational dynamic is encoded while a *p*-coherence order has been selected, the apparent gradient, or the effective gyromagnetic ratio γ, is multiplied by p.[214,215] In this case, the echo signal attenuation $\frac{I}{I_0}$ of the PGSE NMR experiment is governed by a term of the type:[194]

$$e^{\left(-\gamma_{eff}^2\, g^2 D \delta^2\, \Delta'\right)} \tag{3.3}$$

where for homonuclear *p*-quantum coherence, the apparent gyromagnetic ratio $\gamma_{eff} = p\gamma$. **Equation 3.3** is calculated assuming square pulses for the field gradient, and, like in **Equation 1.7**, Δ' is the diffusion delay that takes into account the diffusion during the gradient pulses.

The design of the MQ-DOSY sequence aimed at obtaining a uniform overall diffusion-induced signal loss. In practice, to compensate for the increased sensitivity to magnetic field gradients due to effective gyromagnetic ratio $(p\gamma)$, the gradient pulse duration δ was divided by p.

A crucial aspect for fine separation of mixture components using DOSY experiments is the compensation of convection effects. This is once again demonstrated here. Initially, a standard Hahn spin-echo-based PGSE was integrated into the sequence. In this case, significant convection effects occur and the diffusion-based spin-echo attenuation failed to properly determine diffusion coefficients. In the presence of convection, the spin-echo attenuation is modulated by a cosine term and **Equation 3.3** becomes (**Equation 3.4**):[161,195,219,220]

$$\cos\left(\gamma_{eff}\,\delta g v \Delta\right) e^{\left(-\gamma_{eff}^2\, g^2 D \delta^2\, (\Delta')\right)} \tag{3.4}$$

where v represents the convection velocity. The other parameters are defined in **Equation 3.3**.

As shown in **Equation 3.4**, convection effects appear amplified by the effective gyromagnetic ratio ($p\gamma$), and become a real nuisance, especially when high-quantum orders (4Q in this study) are selected. The incorporation of a convection compensation scheme in 2D and 3D sequences allows recovering characteristic diffusion attenuation for measuring diffusion coefficients.[198,221] The convection compensation effect of the used scheme in **Figure 92** can be easily demonstrated by considering the phase shift at the level of each spin.[222]

After the first gradient echo, the phase shift is given by:

$$\Delta\emptyset_i^{(1)} = \gamma\left(p \times (G_3)\right)\delta z_{0,i} + \gamma\left(p \times (-G_3)\right)\delta\left(z_{0,i} + v_i\,^\Delta/_2\right) = \gamma(-pG_3)\delta\,v_i\,^\Delta/_2$$

where p represents the coherence order, v_i the convection velocity, Δ is always the diffusion delay, $z_{0,i}$ stands for the initial position of spin i, G_3 and δ stand for the strength and the duration of the gradient, respectively.

The π pulse results in the inversion of the sign of the coherences and allows refocusing the chemical shift evolution during the diffusion delay.

$$\Delta\emptyset_i^{(1)} = \gamma(pG_3)\delta\,v_i\,^\Delta/_2$$

The second gradient echo considered independently gives the net phase shift:

$$\Delta\emptyset_i^{(2)} = \gamma\left(p \times (G_3)\right)\delta z_{0,i} + \gamma\left(p \times (-G_3)\right)\delta\left(z_{0,i} + v_i\Delta\right) = \gamma(-pG_3)\delta\,v_i\,^\Delta/_2$$

Thus, at the end of the sequence the sum of the two echoes leads to the complete suppression of convection effects:

$$\Delta\emptyset_i^{(1)} + \Delta\emptyset_i^{(2)} = 0$$

The three-dimensional 4Q-DOSY pulse sequence and the obtained spectrum for the first system, with its three 2D projections, are reported in **Figure 92** and **Figure 93**, respectively.
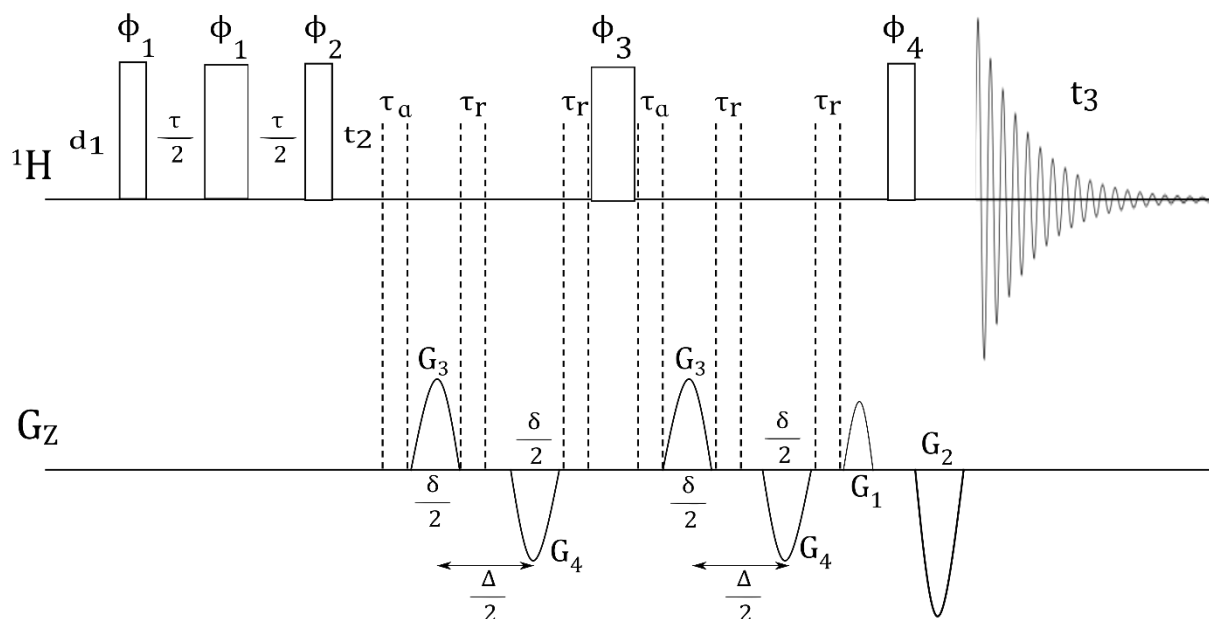
Figure 92 *Pulse Sequences for 3D MaxQ-DOSY. Black and white pulses represent rf pulse flip angles of $\pi/2$ and π, respectively. The phases of the first three pulses where along the x axis. The phase ϕ_2 was chosen to select odd or even MQ orders (x or y for even or odd order excitation, respectively). The duration of the preparation period d_2 was optimized for the uniform excitation of homonuclear MQ of the desired coherence order, the ratio of the gradient pulses was selected to fulfill the intensity of $G_2 = p \times G_1$, where p is the MQ coherence of choice. τ_r means gradient recovery times and τ_a additional delay equal to τ_r introduced to center the second π pulse. For 3D experiment: $\delta = 0.5ms$, $\Delta = 100ms$.*

To resume, 4Q-DOSY means that the diffusion effect was encoded while the 4Q coherence order was excited, which is also the MaxQ order for the six molecules within the reaction mixture. To compensate the increased sensitivity to magnetic field gradients because of effective gyromagnetic ratios ($\gamma_{eff}=p\gamma$), the gradient pulse duration δ was divided by p (here $\delta/4$). A better resolution is provided especially in the 4QDOSY-4Q projection plane with only one spot par molecule located at the crossing of the value of the translational molecular diffusion and the MaxQ (4Q) chemical shift, δ_{MaxQ}. We can hence extract the diffusion coefficient and classify the molecules of the first system according to their affinity to the enzymatic complex (by increasing affinity: 5 and 6 not observed due to the lower sensitivity, and 1 / 3 / 2 / 4). The duration of 3D spectra is a drawback. For this reason, we collect this experiment with the NUS using a compression factor of 4 (25 %), leading to an experiment time of 3 hours (**Figure 93**).
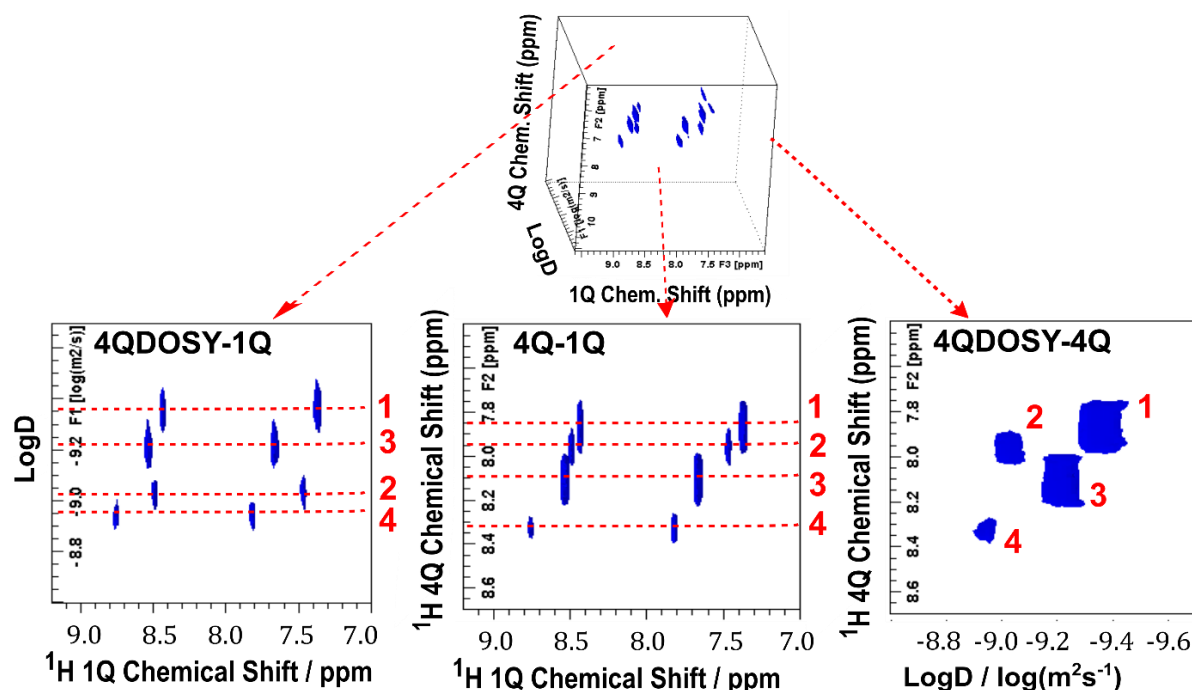
Figure 93 *Projections of the zoom of the 4QDOSY-4Q 3D experiment in the region of 9.5-7 ppm for the first enzyme system. Left: 4QDOSY-1Q projection; middle: 4Q-1Q projection; right: 4QDOSY-4Q projection. The scale in the 4Q dimension, δ'MQ, corresponds to a normalization to the p-quantum order (a division by 4 in this case).*

III.2.7 Conclusions

In this section, we showed that the NMR spectra of mixtures resulting from the reaction of enzyme complexes can be simplified by the well-established 2D Maximum Quantum NMR approach, which allows rapid counting of molecules by isolating the peaks belonging to the same component. For quantification purposes, the simple approach that can be envisaged is to return to the 1D proton spectra to perform the integration of the identified peaks. In the case of mixtures of free small molecules, the delay of the preparation period, τ, of the Multiple Quantum pumping block of 500 ms is accessible. This parameter becomes very limiting when compounds are interacting with large entities and even more when paramagnetic elements are present. Depending on the studied system, the reduction of the relaxation times implies an adjustment to find a compromise between a uniform excitation and loss of the signal. Recent developments with the NUS recording approach enable the acquisition of 2D MQ spectra with short experimental times appropriate to fragile and unstable systems.[159,223] Finally, the

association of the MaxQ and diffusion spectroscopy represents an interesting combination that

provides an extreme simplification of NMR spectra and allows access to valuable information

as the molecular interactions and probably to other physico-chemical information.

General Conclusion and Perspectives

The main work in this thesis was to evaluate the BSS algorithms for the decomposition of NMR spectra of complex mixtures.[126] These algorithms were already used for NMR and showed great potential, calling for further development, to produce rapid and stable algorithms, which can be applied to multidimensional NMR spectra with reasonable computation times, two or three hours, depending on the number of spectra and the power of the computers.   During this thesis, several algorithms have been tested on synthetic mixtures of terpenes which present complex and extremely overlapped 1D spectra. We showed that STALS algorithm is the most adapted for the decomposition of 1D simulated spectra, BC-VMFB for 1D real spectra, while PALM and BC-VMFB-wavelets are the most efficient for the decomposition of the 2D HSQC US and NUS, and TOCSY. For each case, the spectral fingerprint is retrieved for the compounds of the mixtures, despite some spurious peaks. An interesting point in this thesis is the use of SIR, SDR, and Amari indexes on NMR data. These parameters appear to be suitable for evaluating the decomposition of spectra and for estimating the concentrations of components in mixtures. At first sight, SIR can be considered as the most relevant criterion to evaluate separation capabilities of the BSS. However, SIR seems to be blind to the shifts, while SDR goes in the sense of the visual evaluation and it seems more appropriate. There is a need to better evaluate the SIR and SDR indexes for a quantitative assessment of the BSS. Smoothing the spectra with pre-processing before applying the BSS to broaden the peaks can be an interesting perspective for making SIR less sensitive to the shifts. The BSS requires the knowledge of the number of the sources to fit. A statistical analysis, as the Principal Component Analysis, can be helpful for the initialization.

This work highlighted obvious issues concerning the need for pre-processing steps, which were included in the algorithms, to take into account the receiver gain and nc_proc factor, for example. The alignment of the mixture spectra is also a fundamental step. However, even after careful alignment, the Linear Instantaneous Mixture is not the ideal model and has to be improved. These improvements could include, for example, the chemical shift variations or the influence of peak width.

In 2D case, the separation is better and the concentrations are well estimated in comparison to the 1D situation, maybe a consequence of the increased dispersion and data sparsity, also if the computational workload increases. In this context, the reduction of data dimensions to process only the regions containing signals, to significantly reduce the workload, is a promising perspective. Furthermore, the application of the BSS on NUS data sets highlights an interesting feature that can be considered as a denoising filter, which restores source spectra without NUS artefacts. The combination of data dimension reduction and NUS leads to the shortening of the overall analysis time (signal acquisition and signal processing).

Obviously, the BSS works for other 2D experiences, but each one has its own features that need an adaptation. Preliminary results for TOCSY are promising, either in standard acquisition (US) or NUS. In this case, the evaluation indexes are not too satisfactory with the SIR and SDR lower and the Amari index one order of magnitude higher than for 1D and 2D HSQC. This can be due to the presence of the diagonal peaks (more intense) in homonuclear 2D experiments. The attenuation and removing of these latter with particular processing can be a solution.

We explored the use of a single sample with spectroscopic disturbances that mime the concentration variations to blindly decompose NMR spectra. We have shown that, for the diffusion experiments, the concomitant variations of the intensities can be detected by the BSS algorithms if the difference in diffusion coefficients is of an order of magnitude. These results

show that it is possible to associate any spectroscopic disturbances with BSS, provided that they produce a concomitant variation.

In the third chapter, several NMR experiments are applied in order to analyse real complex mixtures, such as biopsies and enzyme systems.

- The results obtained on the assessment of the $T_{1\rho}$ filter in metabolomics are very satisfying in terms of overheating of samples and offset effect. Further developments and implementations in metabolomics studies will be expected in the future with a robust evaluation of the impacts on statistical analyses, using different spin-lock schemes.

- NMR spectra of enzymatic reaction medium can be simplified by the 2D Maximum Quantum NMR approach, which allows rapid counting of components by isolating the peaks belonging to the same compound. We show, using two different enzymatic systems that the implementations imply precautions related to the short relaxation times, in particular, for metalloenzymes or paramagnetic catalysts. Under these conditions, the 2D MaxQ experiment can be used qualitatively, for the assignment of the compounds. The association of the MaxQ and diffusion spectroscopy represents an interesting combination, which leads to a 3D chart that greatly improves the resolution and offers an extreme simplification of the spectra while giving valuable indications on the affinity of the enzymes to the different compounds present in the reaction mixture.

Scientific Publications

The work presented in this thesis has given rise to two publications and one data deposit:

1. Cherni, A.; **Piersanti, E**.; Anthoine, S.; Chaux, C.; Shintu, L.; Yemloul, M.;
    Torrésani, B. Challenges in the Decomposition of 2D NMR Spectra of Mixtures of
    Small Molecules. *Faraday Discuss.* **2019**, *218*, 459–480.
    https://doi.org/10.1039/C9FD00014C.[5]

2. **Piersanti, E**.; Rezig, L.; Tranchida, F.; El-Houri, W.; Abagana, S. M.; Campredon,
    M.; Shintu, L.; Yemloul, M. Evaluation of the Rotating-Frame Relaxation ($T_{1\rho}$)
    Filter and Its Application in Metabolomics as an Alternative to the Transverse
    Relaxation (T_2) Filter. *Anal. Chem.* **2021**, *93* (25), 8746–8753.
    https://doi.org/10.1021/acs.analchem.0c05251.[185]

3. Chaux, C., **Piersanti, E**., Shintu, L., Yemloul, M. NMR Spectra (1H and HSQC) of
    terpenes [Data set]. In Faraday Discussions (v1.0, Vol. 218, pp. 459–480). *Zenodo.*
    2021.
    https://doi.org/10.5281/zenodo.4897295.[126]

Poster and Oral Presentations in Conferences

- 2019/05/13-15: Participation: Challenges in analysis of complex natural mixtures - Faraday Discussion - John McIntyre Conference Centre, University of Edinburgh, UK.

- 2019/05/26-29: Poster (English): New strategy combining 3D-DOSY NMR experiments and the Blind Source Separation for the analysis of mixtures. - Groupement d'Etudes de Résonance Magnétique (GERM) - Saint-Pierre D'Oléron, France.

- 2021/03/26: Poster + 3-minute speech (English): Challenges in the decomposition of NMR spectra of small molecule mixtures. Mini Interdisciplinary Student Symposium (MISS) - Université du Québec, Trois-Rivières, Canada (Online).

- 2021/06/15-16: Poster + 2-minute speech (English): Challenges in the decomposition of NMR spectra of small molecule mixtures. Analytical Research Forum 2021 (ARF21) - United Kingdom (Online).

- 2021/07/1-2: 15-minute speech (English): Challenges in the decomposition of NMR spectra of small molecule mixtures. Journées RMN du Grand Sud - Clermont-Ferrand, France.

- 2021/07/5-8: 15-minute promoted talk (English): Challenges in the decomposition of NMR spectra of small molecule mixtures. EUROMAR 2021 - Virtual platform.

Annexes

Annexe.1 Data Sparsity

In signal processing, statistics or machine learning, sparsity (sometimes called also sparseness or parsimony) refers to situations where a vast majority of data is either zero-valued or extremely small. In a sparse matrix most of the elements are zero, and in our case this is referred to no intensity signals, otherwise the matrix is considered dense (if most of the elements are non-zero valued).[224]

During the last years, sparsity has become a popular paradigm in several contexts, emerging in signal de-noising, inverse problems, *etc.*, where it may be good to search for sparse solution, thus reducing the number of unknowns.[57]

ℓ_1 norm and entropy allow to introduce parsimony. In this work, we impose this on *S*-matrix, considering that the spectra are sparse. We added an indicator to have the data bigger than a small value such as epsilon (non-negative). In the entropy there is a logarithm, so we must have positive data anyway, which is not the case with the standard ℓ_1.[225]

Annexe.2 Majorize-Minimization (MM) algorithm

Majorize-Minimization (MM) algorithm substitutes a difficult minimization problem into a simpler optimization problem.[116–118] It can be defined as an iterative optimization method which at each iteration finds an upper bound for the objective function at the current value, and searches for the minimum of this upper bound. The upper bound has to be such that its minimum can be easily computed. In our approach, following seminal work by Lee and Seung,[115,226] the objective function is split as the sum of a concave function and a convex function (**Figure 94**).
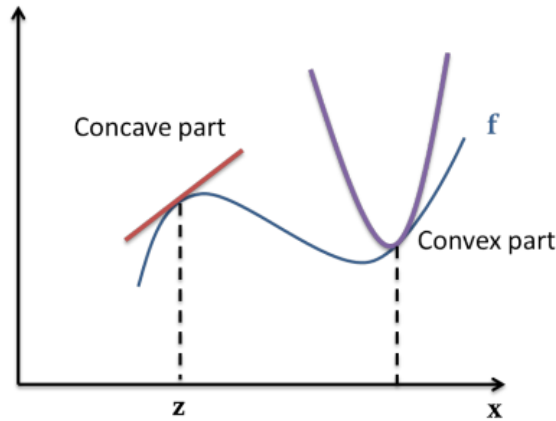
Figure 94 *The given function F(x), which is objective function in our case, is split as the sum of a concave function and a convex function. Here, the example of majorization of a concave function by the tangent. Reprinted by I. Toumi thesis (2013).*

The concave function can be upper bounded by its tangent at the current point, while an upper bound for the convex part can be computed using a technique called Jensen's inequality. The upper bound of the objective function is the sum of the two upper bounds, and a closed form solution can be found in some specific situations (**Figure 95**).
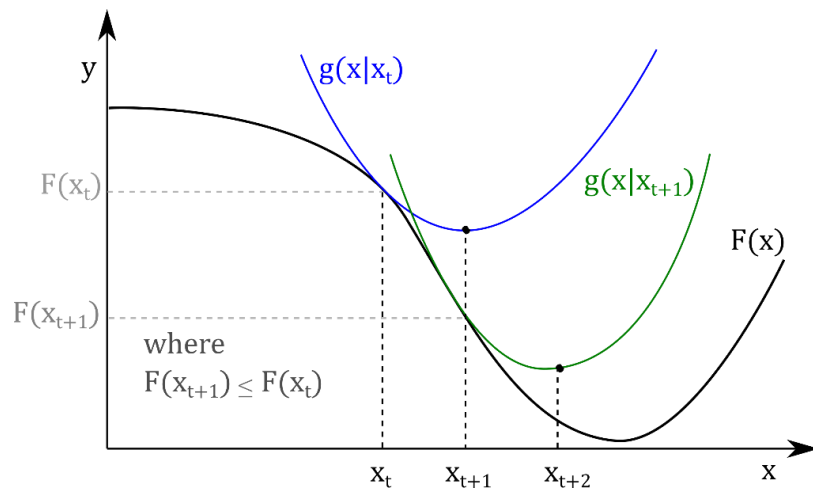
Figure 95 *Explication scheme of Majorize-Minimization (MM) algorithms. It operates by creating one or more auxiliary majorizing functions of a given function* F *to be optimized, starting from a certain point where the tangent passes. Given F(x), the starting point is x_t where a majoring tangent and convex quadratic function is constructed and it is known how to optimize it. The minimum in x_{t+1} is where the calculated gradient is = 0. Another quadratic majorizing tangent function is then constructed in x_{t+1} where the minimum is found in x_{t+2}. The main idea is to optimize the coloured majorizing functions (in blue and green) rather than the more difficult F(x) in black.*

Annexe.3 JADE initialization

Joint Approximate Diagonalization of Eigen-matrices (JADE) is an approach to BSS that exploits a hypothesis of statistical independence of sources. As such it can be called an ICA (Independent Component Analysis) method. ICA generally proceeds by minimizing an objective function that quantifies the degree of dependence between candidate sources. In the case of JADE, the criterion is based on higher order moments. JADE only requires the number of sources to be estimated to start the initialization in function of matrix *X*, with the mixtures of departure, and *K*, the number of the sources (*B=jadeR* (*X, K*)). JADE, using correlations and run on *X*-matrix, produces an estimate for the un-mixing matrix, denoted by *B*. *S=BX* provides an estimate for the sources and the pseudo inverse of *B* (*A=pinvB*) yields an estimate for *A*. Pinv stands for pseudo inverse for the use of the least squares, because inverse matrix may not exist. These estimates in fact take both positive and negative values, thus we use as initialization the absolute values *S=|BX|* and *A=|pinvB|*.

Annexe.4 Projected Alternate Least Squares (PALS) algorithm

PALS algorithm used the most classical objective function with the squared ℓ_2 norm of the difference *X-AS* between data in given *X*-matrix and the ones derived by the LIM model *AS* (**Equation 1.10**):

$$F(X|A,S) = \frac{1}{2} \, ||\, X - AS\, ||\,_F^2$$

The corresponding update rules are given by:

$$Upd_A(A, S) = \Pi_+[(AS - X)S^T\,], \quad Upd_S(A,S) = \Pi_+[A^T(AS-X)],$$

where Π_+ denotes the operator that sets to zero all negative matrix coefficients of its argument.[5]

$(AS - X)S^T$ and $A^T(AS\text{-}X)$ are the solutions of the equations $\nabla F = 0$ with respect to A then with respect to S, where ∇ is the gradient.

The gradient at a point p of a scalar multi-variable differentiable function f is the vector $\nabla f(p)$ which gives the direction in which the function grows fastest in the neighbourhood of p. A gradient exists at every point p where the function f is differentiable, ∇f is called a vector field.

$$\nabla f(\mathrm{p}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathrm{p}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathrm{p}) \end{bmatrix}$$

This is a differentiable optimisation with a F that is differentiable and the gradient can be calculated and annulated.

Annexe.5 Soft Threshold Projected Alternate Least Squares (STALS) algorithm

To enforce the sparsity of the sources, a common practice is to add to the quadratic objective function, an ℓ_1 penalization, which represents the sum of the absolute values of the sources terms (**Equation A.1**):

$$\mathrm{F}(\mathrm{X}|A,S) = \frac{1}{2} ||\,\mathrm{X} - AS||_F^2 + \lambda_s||\mathrm{S}||_1 \qquad (\mathrm{A.1})$$

where $||\mathrm{S}||_1$ is the so-called ℓ_1 norm of *S*, in other words the sum of absolute values of the elements of *S*, and λ_s is a positive constant which modulates the strength of the penalty, setting to zero the values whose absolute values are smaller than the threshold λ_s. It is user defined, and must be carefully chosen. If λ_S is taken too big, part of the signal is erased and considered as noise, while if λ_S is set too low, part of the noise is kept into the signal. In our work we have

systematically compared results obtained with 0,01 σ, 0,1σ, 1σ, 10σ, 100 σ, *etc.*, where σ is the standard deviation of the noise estimated from real data, using a signal-free segment.

To solve this problem, commonly, the projection Π_+ of S is replaced by the non-negative soft thresholding operator S_λ, which sets to zero all matrix coefficients smaller than the threshold. The corresponding update rules become:

$$\text{Upd}_A(A,S) = \Pi_+[(AS - X)S^T], \quad \text{Upd}_S(A,S) = S_\lambda[A^T(AS-X)],$$

where PALS coincide with STALS when $\lambda_S = 0$.

Annexe.6 Proximal Alternating Linearized Minimization (PALM) and Block Coordinate Variable Metric Forward- Backward algorithms

Proximal Alternating Linearized Minimization (PALM) and Block Coordinate Variable Metric Forward- Backward have the objective to minimize the cost function (**Equation A.1**), considering both the minimization of the quadratic part and the regularization part.[227] They are based on a projected gradient descent algorithm. BC-VMFB algorithm is a pre-conditioned algorithm using variable metrics (VM) allowing the increase of convergence speed.[227] It uses the D. D. Lee and H. S. Seung preconditioning, where γ_g, the gradient descent step size, is represented by a preconditioning matrix.[115]

PALM is different from BC-VMFB because has not the preconditioning step and γ_g is scalar. The update rules for BC-VMFB and PALM are given by:

$$\text{Upd}_A(A,S) = \Pi_+[A - \gamma_g(AS - X)S^T],$$

$$\text{Upd}_S(A, S) = (S_{\lambda/\gamma_g}[\psi^T(S - \gamma_g A^T(AS-X))]),$$

The γ_g is used in convex optimization to minimize and find the location of the global minimum of a multi-variable function. Since the gradient $\nabla f(p)$ gives the direction in which the function

grows faster, to find the minimum, γ_g goes in the opposite direction to that of the gradient. The algorithm starts at an initial point, then moving a small distance in the negative of the gradient, will go to a new point more or less quickly thanks to γ_g. The moved distance is in fact decided by the step size γ_g. This new point approaches the solution. If the value of γ_g is too small, the algorithm will be too slow to find the global minimum, while if it is too big we miss the minimum and the algorithm diverges.

In PALM, the fixed step size is $\gamma_g < 1/\kappa$ where κ is the Lipschitz constant, derivation of the Lipschitz continuity.

As shown in **Figure 96**, *given two metric spaces (X, d_x) and (Y, d_y) where d_x denotes the metric on the set X and d_y is the metric on set Y, a function $f: X \rightarrow Y$ is called Lipschitz continuous if there exists a real constant $\kappa \geq 0$ such that, for all x_1 and x_2 in X:*[228]

$$d_y(f(x_1),\ f(x_2)) \leq \kappa\ d_x(x_1,\ x_2)$$

and

$$/f(x_1) - f(x_2)| \leq \kappa\ |x_1 - x_2|$$

where κ is referred to the Lipschitz constant for the function f. The smallest constant is usually the best constant.
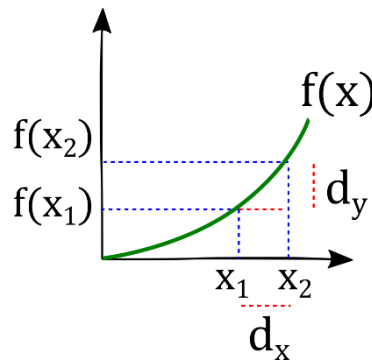
Figure 96 *Explication scheme for the Lipschitz constant.*

$F: \mathbb{R}^N \rightarrow \mathbb{R}$ *is differentiable. Moreover, F has an κ-Lipschitzian gradient where $\kappa > 0$:*[227]

$$||\nabla F(x) - \nabla F(y)|| \leq \kappa\ ||x - y||$$

This is possible only if $f(x)$ is a convex function (**Figure 97**):

$f(x)$ is a convex function if and only if for all real $0 < t < 1$ and all x_1, $x_2 \in X$ such that $x_1 \neq x_2$:

$$f(tx_1 + (1\text{-}t)\,x_1) \leq t\,f(x_1) + (1\text{-}t)\,f(x_2)$$
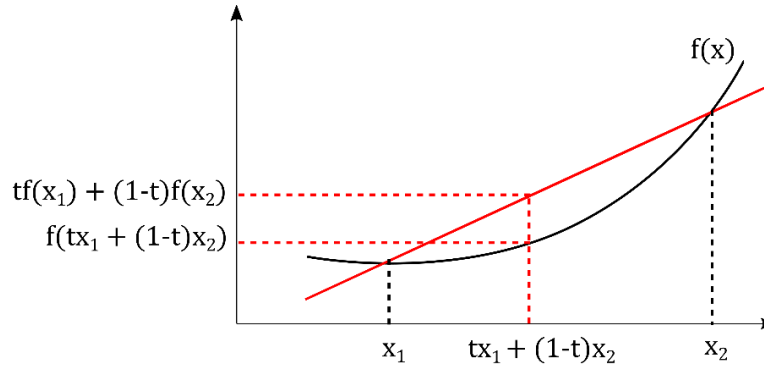
Figure 97 *Example of a convex function.*

Annexe.7 Changing the representation domain using wavelets

The algorithms used also change with respect to the source representation domain, which can be in spectral domain or wavelet domain. This regards a new optimization method, different from the alternate approach on *A* and *S* for the bi-convex problem, used to optimize the least squares criterion, with respect our two variables *A* and *S*.

In digital signal processing, the word wavelet comes from Morlet and Grossmann that used it in the early 1980s. The French translation is "ondelette" meaning small wave.

The use of wavelets to process NMR signals has been exploited by several authors where the main advantage was its ability to compress signals.[229] [230] [231] A wavelet is a wave-like oscillation defined as a sinus function. Using convolution techniques, wavelets can be combined with known portions of a damaged signal, for example, to extract information from the unknown region.

It is a useful mathematical tool that can be applied to several subjects, in different kind of data, but also to extract information from audio signals and images. In data analysis, they are

generally used as sets of complementary wavelets to reversibly decompose data without gaps or overlaps. Wavelet compression/decompression algorithms are useful to recover the original information with minimal loss.

The mixtures and sources are represented by point values, $S_{k\ell}$ and $X_{n\ell}$. The objective functions are separable. For a certain A, columns of S (same regions if spectra for different sources) are processed independently of each other. Possible correlations in the spectral domain (represented by index ℓ) are not exploited. It is possible, in this way, to describe the spectral domain using a different representation based on an expansion on a set of L-dimensional vectors, which form a basis of the L-dimensional space. Using this representation, a regularization is introduced on the corresponding coefficients rather than the source matrix S. If these vectors are noted as $\{\psi^{(c)}, c=1, \dots L\}$, and therefore they are concatenated in a square matrix denoted by ψ (where the columns of ψ are the vectors $\psi^{(c)}$), it may be shown that the coefficients of the source matrix in this basis are given by matrix $\Gamma = \psi^{T}S$ (where ^{T} stands for matrix transposition). This new adaptation of the generic objective function can be re-written as:

$$F(X|A,S) = \frac{1}{2} || X - AS ||_F^2 + f_A(A) + f_\Gamma(\psi^T S),$$

where f_Γ is a penalty function. We will choose an ℓ_1 penalization with a regularization parameter $\lambda \geq 0$. $f_\Gamma(\Gamma) = \lambda \Sigma_{k,l} |\gamma_{k,l}|$, where the coefficients of matrix Γ are $\gamma_{k,l}$ in the optimization in the wavelet transform domain. This enforces the sparsity, concentrating the information in very few coefficients, with a larger number of coefficients close or equal to zero.[5]

Annexe.8 Wavelet-based PALM and BC-VMFB Algorithm

Here the sparsity is imposed on the wavelet coefficients of the spectra $\gamma_{m,l}$. The objective function will be:

$$F(X|A,S) = \frac{1}{2} \, || \, X - AS || \, _F^2 \, + \lambda ||\psi^T S||_1,$$

where $||\Gamma||_1$ $(\Gamma = \psi^T S)$ is the sum of the absolute values of coefficients $\gamma_{m,l}$. The variation

from BC-VMFB and PALM is that here the threshold operation is done on wavelets

coefficients rather than spectrum coefficients.

$$Upd_A(A,S) = \Pi_+[A - \gamma_g \, (AS - X)S^T \,],$$

$$Upd_S(A, S) = \psi(S_{\lambda/} \, \gamma_g \, [\psi^T(S - \gamma_g \, A^T(AS-X))]),$$

where S_λ sets only the values whose absolute value is smaller than λ_S to zero.[232]

Bibliographic References

(1)   Chaux, C.; Combettes, P. L.; Pesquet, J.-C.; Wajs, V. R. A Variational Formulation for Frame-Based Inverse Problems. *Inverse Probl.* **2007**, *23* (4), 1495–1518. https://doi.org/10.1088/0266-5611/23/4/008.

(2)   Belouchrani, A.; Abed-Meraim, K.; Cardoso, J.-F.; Moulines, E. A Blind Source Separation Technique Using Second-Order Statistics. *IEEE Trans. Signal Process.* **1997**, *45* (2), 434–444. https://doi.org/10.1109/78.554307.

(3)   Jutten, C.; Herault, J. Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture. *Signal Process.* **1991**, *24* (1), 1–10. https://doi.org/10.1016/0165-1684(91)90079-X.

(4)   Toumi, I.; Caldarelli, S.; Torrésani, B. A Review of Blind Source Separation in NMR Spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *81*, 37–64. https://doi.org/10.1016/j.pnmrs.2014.06.002.

(5)   Cherni, A.; Piersanti, E.; Anthoine, S.; Chaux, C.; Shintu, L.; Yemloul, M.; Torrésani, B. Challenges in the Decomposition of 2D NMR Spectra of Mixtures of Small Molecules. *Faraday Discuss.* **2019**, *218*, 459–480. https://doi.org/10.1039/C9FD00014C.

(6)   Cherni, A.; Piersanti, E.; Chaux, C. NMF-Based Sparse Unmixing of Complex Mixtures. 2.

(7)   Keeler, J. *Understanding NMR Spectroscopy*, second edition.; Wiley, 2010.

(8)   Lin, M.; Shapiro, M. J. Mixture Analysis by NMR Spectroscopy. *Anal. Chem.* **1997**, *69* (22), 4731–4733. https://doi.org/10.1021/ac970594x.

(9)   Kalic, M. Assignment of Overlapping 1H NMR Signals in Carp Seminal Plasma by Proton-Detected 2D C,H Correlation Spectroscopy. *Biochem. Biophys. Res. Commun.* **2000**, *269* (1), 5.

(10)   Rastogi, V. K. 31P–1H Correlation and Resonance Assignments along the DNA Backbone: Three-dimensional Implementation of Heteronuclear Long-range Correlation Experiment. *Magn Reson Chem* **2000**, 5.

(11)   Xiao, H.-H.; Lv, J.; Mok, D.; Yao, X.-S.; Wong, M.-S.; Cooper, R. NMR Applications for Botanical Mixtures: The Use of HSQC Data to Determine Lignan Content in *Sambucus Williamsii*. *J. Nat. Prod.* **2019**, *82* (7), 1733–1740. https://doi.org/10.1021/acs.jnatprod.8b00891.

(12)   Morris, G. A.; Freeman, R. Enhancement of Nuclear Magnetic Resonance Signals by Polarization Transfer. *J. Am. Chem. Soc.* **1979**, 3.

(13)   Cavanagh, J.; Palmer, A. G.; Wright, P. E.; Rance, M. Sensitivity Improvement in Proton-Detected Two-Dimensional Heteronuclear Relay Spectroscopy. *J. Magn. Reson. 1969* **1991**, *91* (2), 429–436. https://doi.org/10.1016/0022-2364(91)90209-C.

(14)   Farjon, J.; Milande, C.; Martineau, E.; Akoka, S.; Giraudeau, P. The FAQUIRE Approach: FAst, QUantitative, HIghly Resolved and SEnsitivity Enhanced [1] H, [13] C Data. *Anal. Chem.* **2018**, *90* (3), 1845–1851. https://doi.org/10.1021/acs.analchem.7b03874.

(15)   Heikkinen, S.; Toikka, M. M.; Karhunen, P. T.; Kilpeläinen, I. A. Quantitative 2D HSQC (Q-HSQC) via Suppression of *J* -Dependence of Polarization Transfer in NMR Spectroscopy: Application to Wood Lignin. *J. Am. Chem. Soc.* **2003**, *125* (14), 4362–4367. https://doi.org/10.1021/ja029035k.

(16)   Peterson, D. J.; Loening, N. M. QQ-HSQC: A Quick, Quantitative Heteronuclear Correlation Experiment for NMR Spectroscopy. *Magn. Reson. Chem.* **2007**, *45* (11), 937–941. https://doi.org/10.1002/mrc.2073.

(17)   Koskela, H.; Kilpeläinen, I.; Heikkinen, S. Some Aspects of Quantitative 2D NMR. *J. Magn. Reson.* **2005**, *174* (2), 237–244. https://doi.org/10.1016/j.jmr.2005.02.002.

(18)   Koskela, H.; Heikkilä, O.; Kilpeläinen, I.; Heikkinen, S. Quantitative Two-Dimensional HSQC Experiment for High Magnetic Field NMR Spectrometers. *J. Magn. Reson.* **2010**, *202* (1), 24–33. https://doi.org/10.1016/j.jmr.2009.09.021.

(19) Frydman, L.; Scherf, T.; Lupulescu, A. The Acquisition of Multidimensional NMR Spectra within a Single Scan. *Proc. Natl. Acad. Sci.* **2002**, *99* (25), 15858–15862. https://doi.org/10.1073/pnas.252644399.

(20) Schanda, P.; Kupče, Ē.; Brutscher, B. SOFAST-HMQC Experiments for Recording Two-Dimensional Deteronuclear Correlation Spectra of Proteins within a Few Seconds. *J. Biomol. NMR* **2005**, *33* (4), 199–211. https://doi.org/10.1007/s10858-005-4425-x.

(21) Schanda, P.; Brutscher, B. Very Fast Two-Dimensional NMR Spectroscopy for Real-Time Investigation of Dynamic Events in Proteins on the Time Scale of Seconds. *J. Am. Chem. Soc.* **2005**, *127* (22), 8014–8015. https://doi.org/10.1021/ja051306e.

(22) Maciejewski, M. W.; Mobli, M.; Schuyler, A. D.; Stern, A. S.; Hoch, J. C. Data Sampling in Multidimensional NMR: Fundamentals and Strategies. In *Novel Sampling Approaches in Higher Dimensional NMR*; Billeter, M., Orekhov, V., Eds.; Topics in Current Chemistry; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; Vol. 316, pp 49–77. https://doi.org/10.1007/128_2011_185.

(23) Mobli, M. Nonuniform Sampling and Non-Fourier Signal Processing Methods in Multidimensional NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, 23.

(24) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. Improved Resolution in Triple-Resonance Spectra by Nonlinear Sampling in the Constant-Time Domain. *J. Biomol. NMR* **1994**, *4* (4), 483–490. https://doi.org/10.1007/BF00156615.

(25) Rovnyak, D.; Frueh, D. P.; Sastry, M.; Sun, Z.-Y. J.; Stern, A. S.; Hoch, J. C.; Wagner, G. Accelerated Acquisition of High Resolution Triple-Resonance Spectra Using Non-Uniform Sampling and Maximum Entropy Reconstruction. *J. Magn. Reson.* **2004**, *170* (1), 15–21. https://doi.org/10.1016/j.jmr.2004.05.016.

(26) Frueh, D. P.; Sun, Z.-Y. J.; Vosburg, D. A.; Walsh, C. T.; Hoch, J. C.; Wagner, G. Non-Uniformly Sampled Double-TROSY HNcaNH Experiments for NMR Sequential Assignments of Large Proteins. *J. Am. Chem. Soc.* **2006**, *128* (17), 5757–5763. https://doi.org/10.1021/ja0584222.

(27) Mobli, M.; Maciejewski, M. W.; Gryk, M. R.; Hoch, J. C. Automatic Maximum Entropy Spectral Reconstruction in NMR. *J. Biomol. NMR* **2007**, *39* (2), 133–139. https://doi.org/10.1007/s10858-007-9180-8.

(28) Orekhov, V. Y.; Ibraghimov, I.; Billeter, M. Optimizing Resolution in Multidimensional NMR by Three-Way Decomposition. 9.

(29) Jaravine, V.; Ibraghimov, I.; Yu Orekhov, V. Removal of a Time Barrier for High-Resolution Multidimensional NMR Spectroscopy. *Nat. Methods* **2006**, *3* (8), 605–607. https://doi.org/10.1038/nmeth900.

(30) Luan, T.; Jaravine, V.; Yee, A.; Arrowsmith, C. H.; Orekhov, V. Yu. Optimization of Resolution and Sensitivity of 4D NOESY Using Multi-Dimensional Decomposition. *J. Biomol. NMR* **2005**, *33* (1), 1–14. https://doi.org/10.1007/s10858-005-1363-6.

(31) Kazimierczuk, K.; Orekhov, V. Yu. Accelerated NMR Spectroscopy by Using Compressed Sensing. *Angew. Chem. Int. Ed.* **2011**, *50* (24), 5556–5559. https://doi.org/10.1002/anie.201100370.

(32) Kazimierczuk, K.; Zawadzka, A.; Koźmiński, W.; Zhukov, I. Random Sampling of Evolution Time Space and Fourier Transform Processing. *J. Biomol. NMR* **2006**, *36* (3), 157–168. https://doi.org/10.1007/s10858-006-9077-y.

(33) Kazimierczuk, K.; Koźmiński, W.; Zhukov, I. Two-Dimensional Fourier Transform of Arbitrarily Sampled NMR Data Sets. *J. Magn. Reson.* **2006**, *179* (2), 323–328. https://doi.org/10.1016/j.jmr.2006.02.001.

(34) Kazimierczuk, K.; Zawadzka, A.; Koźmiński, W. Optimization of Random Time Domain Sampling in Multidimensional NMR. *J. Magn. Reson.* **2008**, *192* (1), 123–130. https://doi.org/10.1016/j.jmr.2008.02.003.

(35) Guennec, A. L.; Giraudeau, P.; Caldarelli, S. Evaluation of Fast 2D NMR for Metabolomics. *Anal. Chem.* **2014**, *86* (12), 5946–5954. https://doi.org/10.1021/ac500966e.

(36) Aoto, P. C.; Fenwick, R. B.; Kroon, G. J. A.; Wright, P. E. Accurate Scoring of Non-Uniform Sampling Schemes for Quantitative NMR. *J. Magn. Reson.* **2014**, *246*, 31–35. https://doi.org/10.1016/j.jmr.2014.06.020.

(37)   Bloch, F. Dynamical Theory of Nuclear Induction. II. *Phys. Rev.* **1956**, *102* (1), 104–135. https://doi.org/10.1103/PhysRev.102.104.

(38)   Wangsness, R. K.; Bloch, F. The Dynamical Theory of Nuclear Induction. *Phys. Rev.* **1953**, *89* (4), 728–739. https://doi.org/10.1103/PhysRev.89.728.

(39)   Redfield, A. G. On the Theory of Relaxation Processes. *IBM J. Res. Dev.* **1957**, *1* (1), 19–31. https://doi.org/10.1147/rd.11.0019.

(40)   Vold, R. L.; Waugh, J. S.; Klein, M. P.; Phelps, D. E. Measurement of Spin Relaxation in Complex Systems. *J. Chem. Phys.* **1968**, *48* (8), 3831–3832. https://doi.org/10.1063/1.1669699.

(41)   Carr, H. Y.; Purcell, E. M. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. *Phys. Rev.* **1954**, *94* (3), 630–638. https://doi.org/10.1103/PhysRev.94.630.

(42)   Hahn, E. L. Spin Echoes. *Phys. Rev.* **1950**, *80* (4), 580–594. https://doi.org/10.1103/PhysRev.80.580.

(43)   Takegoshi, K.; Ogura, K.; Hikichi, K. A Perfect Spin Echo in a Weakly Homonuclear J-Coupled Two Spin- System. *J. Magn. Reson. 1969* **1989**, *84* (3), 611–615. https://doi.org/10.1016/0022-2364(89)90127-3.

(44)   Aguilar, J. A.; Nilsson, M.; Bodenhausen, G.; Morris, G. A. Spin Echo NMR Spectra without J Modulation. *Chem Commun* **2012**, *48* (6), 811–813. https://doi.org/10.1039/C1CC16699A.

(45)   Tourell, M. C.; Pop, I.-A.; Brown, L. J.; Brown, R. C. D.; Pileio, G. Singlet-Assisted Diffusion-NMR (SAD-NMR): Redefining the Limits When Measuring Tortuosity in Porous Media. *Phys. Chem. Chem. Phys.* **2018**, *20* (20), 13705–13713. https://doi.org/10.1039/C8CP00145F.

(46)   Stilbs, Peter. Molecular Self-Diffusion Coefficients in Fourier Transform Nuclear Magnetic Resonance Spectrometric Analysis of Complex Mixtures. *Anal. Chem.* **1981**, *53* (13), 2135–2137. https://doi.org/10.1021/ac00236a044.

(47)   Barjat, H.; Gilles, G. A.; Smart, S.; Swanson, A. G.; Williams, S. C. R. High-Resolution Diffusion-Ordered 2D Spectroscopy (HR-DOSY)—a New Tool for the Analysis of Complex Mixtures. *J. Magn. Reson.* 1995, pp 170–172.

(48)   Morris, K. F.; Stilbs, Peter.; Johnson, C. S. Analysis of Mixtures Based on Molecular Size and Hydrophobicity by Means of Diffusion-Ordered 2D NMR. *Anal. Chem.* **1994**, *66* (2), 211–215. https://doi.org/10.1021/ac00074a006.

(49)   Nilsson, M.; Duarte, I. F.; Almeida, C.; Delgadillo, I.; Goodfellow, B. J.; Gil, A. M.; Morris, G. A. High-Resolution NMR and Diffusion-Ordered Spectroscopy of Port Wine. *J. Agric. Food Chem.* **2004**, *52* (12), 3736–3743. https://doi.org/10.1021/jf049797u.

(50)   Nilsson, M. The DOSY Toolbox: A New Tool for Processing PFG NMR Diffusion Data. *J. Magn. Reson.* **2009**, *200* (2), 296–302. https://doi.org/10.1016/j.jmr.2009.07.022.

(51)   Stejskal, E. O.; Tanner, J. E. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *J. Chem. Phys.* **1965**, *42* (1), 288–292. https://doi.org/10.1063/1.1695690.

(52)   Tanner, J. E. Use of the Stimulated Echo in NMR Diffusion Studies. *J. Chem. Phys.* **1970**, *52* (5), 2523–2526. https://doi.org/10.1063/1.1673336.

(53)   Stilbs, P. Fourier Transform Pulsed-Gradient Spin-Echo Studies of Molecular Diffusion. *Prog. Nucl. Magn. Reson. Spectrosc.* **1987**, *19* (1), 1–45. https://doi.org/10.1016/0079-6565(87)80007-9.

(54)   Morris, K. F.; Johnson, C. S. Diffusion-Ordered Two-Dimensional Nuclear Magnetic Resonance Spectroscopy. *J. Am. Chem. Soc.* **1992**, *114* (8), 3139–3141. https://doi.org/10.1021/ja00034a071.

(55)   Cohen, Y.; Avram, L.; Frish, L. Diffusion NMR Spectroscopy in Supramolecular and Combinatorial Chemistry: An Old Parameter?New Insights. *Angew. Chem. Int. Ed.* **2005**, *44* (4), 520–554. https://doi.org/10.1002/anie.200300637.

(56)   Bryan, K. Elementary Inversion of the Laplace Transform. 10.

(57)   Toumi, I. *Decomposition Methods of NMR Signal of Complex Mixtures: Models and Applications*; Aix Marseille Université, 2013; Vol. Thèse de Doctorat.

(58)   Evans, R.; Day, I. J. Matrix-Assisted Diffusion-Ordered Spectroscopy. *RSC Adv.* **2016**, *6* (52), 47010–47022. https://doi.org/10.1039/C6RA05380G.

(59) Spraul, M.; Humpfer, E.; Sch, H. NMR-BASED MIXTURE ANALYSIS ON THE EXAMPLE OF FRUIT JUICE QUALITY CONTROL USING STATISTICS AND QUANTIFICATION. 22.

(60) Buedenbender, L.; Habener, L. J.; Grkovic, T.; Kurtböke, D. İ.; Duffy, S.; Avery, V. M.; Carroll, A. R. HSQC–TOCSY Fingerprinting for Prioritization of Polyketide- and Peptide-Producing Microbial Isolates. *J. Nat. Prod.* **2018**, *81* (4), 957–965. https://doi.org/10.1021/acs.jnatprod.7b01063.

(61) Barjat, H.; Morris, G. A.; Swanson, A. G. A Three-Dimensional DOSY–HMQC Experiment for the High-Resolution Analysis of Complex Mixtures. *J. Magn. Reson.* **1998**, *131* (1), 131–138. https://doi.org/10.1006/jmre.1997.1332.

(62) Reddy, G. N. M.; Caldarelli, S. Maximum-Quantum (MaxQ) NMR for the Speciation of Mixtures of Phenolic Moleculeswz. **2011**, 3.

(63) Manjunatha Reddy, G. N.; Yemloul, M.; Caldarelli, S. Combined Maximum-Quantum and DOSY 3D Experiments Provide Enhanced Resolution for Small Molecules in Mixtures: MaxQ-MaxQDOSY. *Magn. Reson. Chem.* **2017**, *55* (5), 492–497. https://doi.org/10.1002/mrc.4465.

(64) Nilsson, M.; Morris, G. A. Pure Shift Proton DOSY: Diffusion-Ordered 1H Spectra without Multiplet Structure. *Chem. Commun.* **2007**, No. 9, 933. https://doi.org/10.1039/b617761a.

(65) Foroozandeh, M.; Morris, G. A.; Nilsson, M. PSYCHE Pure Shift NMR Spectroscopy. *Chem. – Eur. J.* **2018**, *24* (53), 13988–14000. https://doi.org/10.1002/chem.201800524.

(66) Zangger, K. Pure Shift NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **2015**, *86–87*, 1–20. https://doi.org/10.1016/j.pnmrs.2015.02.002.

(67) Petersen, B. O.; Nilsson, M.; Bøjstrup, M.; Hindsgaul, O.; Meier, S. 1H NMR Spectroscopy for Profiling Complex Carbohydrate Mixtures in Non-Fractionated Beer. *Food Chem.* **2014**, *150*, 65–72. https://doi.org/10.1016/j.foodchem.2013.10.136.

(68) Nilsson, M.; Morris, G. A. Speedy Component Resolution: An Improved Tool for Processing Diffusion-Ordered Spectroscopy Data. *Anal. Chem.* **2008**, *80* (10), 3777–3782. https://doi.org/10.1021/ac7025833.

(69) Manjunatha Reddy. *NMR Methods and Applications for the Study of Complex Mixtures*, Thèse de Doctorat.; Aix Marseille Université, 2012.

(70) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

(71) Nilsson, M. The DOSY Toolbox: A New Tool for Processing PFG NMR Diffusion Data. *J. Magn. Reson.* **2009**, *200* (2), 296–302. https://doi.org/10.1016/j.jmr.2009.07.022.

(72) Castañar, L.; Poggetto, G. D.; Colbourne, A. A.; Morris, G. A.; Nilsson, M. The GNAT: A New Tool for Processing NMR Data. *Magn. Reson. Chem.* **2018**, *56* (6), 546–558. https://doi.org/10.1002/mrc.4717.

(73) Cherni, A.; Chouzenoux, E.; Delsuc, M.-A. PALMA, an Improved Algorithm for DOSY Signal Processing. *The Analyst* **2017**, *142* (5), 772–779. https://doi.org/10.1039/C6AN01902A.

(74) Colbourne, A. A.; Meier, S.; Morris, G. A.; Nilsson, M. Unmixing the NMR Spectra of Similar Species – Vive La Diffe´rence. **2013**, 3.

(75) Colbourne, A. A.; Meier, S.; Morris, G. A.; Nilsson, M. Unmixing the NMR Spectra of Similar Species – Vive La Différence. *Chem. Commun.* **2013**, *49* (89), 10510. https://doi.org/10.1039/c3cc46228e.

(76) Windig, W.; Antalek, B. Direct Exponential Curve Resolution Algorithm (DECRA): A Novel Application of the Generalized Rank Annihilation Method for a Single Spectral Mixture Data Set with Exponentially Decaying Contribution Profiles. *Chemom. Intell. Lab. Syst.* **1997**, *37* (2), 241–254. https://doi.org/10.1016/S0169-7439(97)00028-2.

(77) Nilsson, M.; Morris, G. A. Improved DECRA Processing of DOSY Data: Correcting for Non-Uniform Field Gradients. *Magn. Reson. Chem.* **2007**, *45* (8), 656–660. https://doi.org/10.1002/mrc.2023.

(78) Morris, G. A.; Barjat, H.; Home, T. J. Reference Deconvolution Methods. *Prog. Nucl. Magn. Reson. Spectrosc.* **1997**, *31* (2–3), 197–257. https://doi.org/10.1016/S0079-6565(97)00011-3.

(79) Harshman, R. A. Parafac. UCLA Working Papers in Phonetics. 1970, p 1.

(80) Harshman, R. A.; Lundy, M. E. PARAFAC: Parallel Factor Analysis. *Comput. Stat. Data Anal.* **1994**, *18* (1), 39–72. https://doi.org/10.1016/0167-9473(94)90132-5.

(81) Tønning, E.; Polders, D.; Callaghan, P. T.; Engelsen, S. B. A Novel Improved Method for Analysis of 2D Diffusion–Relaxation Data—2D PARAFAC-Laplace Decomposition. *J. Magn. Reson.* **2007**, *188* (1), 10–23. https://doi.org/10.1016/j.jmr.2007.05.018.

(82) Nilsson, M.; Khajeh, M.; Botana, A.; Bernstein, M. A.; Morris, G. A. Diffusion NMR and Trilinear Analysis in the Study of Reaction Kinetics. *Chem. Commun.* **2009**, No. 10, 1252. https://doi.org/10.1039/b820813a.

(83) Jensen, K. S.; Linse, S.; Nilsson, M.; Akke, M.; Malmendal, A. Revealing Well-Defined Soluble States during Amyloid Fibril Formation by Multilinear Analysis of NMR Diffusion Data. *J. Am. Chem. Soc.* **2019**, *141* (47), 18649–18652. https://doi.org/10.1021/jacs.9b07952.

(84) Delsuc, M. A.; Malliavin, T. E. Maximum Entropy Processing of DOSY NMR Spectra. *Anal. Chem.* **1998**, *70* (10), 2146–2148. https://doi.org/10.1021/ac9800715.

(85) Vincent, E.; Gribonval, R.; Fevotte, C. Performance Measurement in Blind Audio Source Separation. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14* (4), 1462–1469. https://doi.org/10.1109/TSA.2005.858005.

(86) Isomura, T.; Toyoizumi, T. Multi-Context Blind Source Separation by Error-Gated Hebbian Rule. *Sci. Rep.* **2019**, *9* (1), 7127. https://doi.org/10.1038/s41598-019-43423-z.

(87) Cichocki, A.; Amari, S.-I. *Adaptative Blind Signal and Image Processing: Learning Algorithms and Applications.*, John Wiley&Sons, Inc.; New York, NY, USA., 2002.

(88) Belouchrani, A.; Amin, M. G. Blind Source Separation Based on Time-Frequency Signal Representations. *IEEE Trans. Signal Process.* **1998**, *46* (11), 2888–2897. https://doi.org/10.1109/78.726803.

(89) Amin, M. G.; Belouchrani, A. Blind Source Separation Using the Spatial Ambiguity Functions. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380)*; IEEE: Pittsburgh, PA, USA, 1998; pp 413–416. https://doi.org/10.1109/TFSA.1998.721449.

(90) Comon, P. Blind Channel Identification and Extraction of More Sources than Sensors; Luk, F. T., Ed.; San Diego, CA, USA, 1998; p 2. https://doi.org/10.1117/12.325670.

(91) Comon, P. Independent Component Analysis, A New Concept? *Signal Process.* **1994**, *36* (3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9.

(92) Macchi, O.; Moreau, E. HIGH-ORDER CONTRASTS FOR SELF-ADAPTIVE SOURCE SEPARATION. 28.

(93) Emile, B.; Comon, P.; Le Roux, J. Estimation of Time Delays with Fewer Sensors than Sources. *IEEE Trans. Signal Process.* **1998**, *46* (7), 2012–2015. https://doi.org/10.1109/78.700972.

(94) Boustany, R.; Antoni, J. A Subspace Method for the Blind Extraction of a Cyclostationary Source: Application to Rolling Element Bearing Diagnostics. *Mech. Syst. Signal Process.* **2005**, *19* (6), 1245–1259. https://doi.org/10.1016/j.ymssp.2005.08.016.

(95) Yong Xiang; Abed-Meraim, K.; Yingbo Hua. Adaptive Blind Source Separation by Second Order Statistics and Natural Gradient. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*; IEEE: Phoenix, AZ, USA, 1999; pp 2917–2920 vol.5. https://doi.org/10.1109/ICASSP.1999.761373.

(96) Leyman, A. R.; Kamran, Z. M.; Abed-Meraim, K. Higher-Order Time Frequency-Based Blind Source Separation Technique. *IEEE Signal Process. Lett.* **2000**, *7* (7), 193–196. https://doi.org/10.1109/97.847366.

(97) Lahat, D.; Cardoso, J.-F.; Messer, H. Second-Order Multidimensional ICA: Performance Analysis. *IEEE Trans. Signal Process.* **2012**, *60* (9), 4598–4610. https://doi.org/10.1109/TSP.2012.2199985.

(98) Tanaka, A.; Imai, H.; Miyakoshi, M. Second-Order-Statistics-Based Blind Source Separation for Non-Stationary Sources with Stationary Noise. In *2006 IEEE International Symposium on Signal Processing and Information Technology*; IEEE: Vancouver, BC, Canada, 2006; pp 647–650. https://doi.org/10.1109/ISSPIT.2006.270880.

(99) Tanaka, A.; Imai, H.; Miyakoshi, M. Selection of Correlation Matrices for Second-Order-Statistics-Based Blind Source Separation. In *2007 IEEE/SP 14th Workshop on Statistical Signal*

Processing; IEEE: Madison, WI, USA, 2007; pp 109–113. https://doi.org/10.1109/SSP.2007.4301228.

(100) Puntonet, C. G.; Prieto, A.; Jutten, C.; Rodríguez-Alvarez, M.; Ortega, J. Separation of Sources: A Geometry-Based Procedure for Reconstruction of n-Valued Signals. *Signal Process.* **1995**, *46* (3), 267–284. https://doi.org/10.1016/0165-1684(95)00088-0.

(101) Kamran, Z. M.; Leyman, A. R.; Abed-Meraim, K. Techniques for Blind Source Separation Using Higher-Order Statistics. In *Proceedings of the Tenth IEEE Workshop on Statistical Signal and Array Processing (Cat. No.00TH8496)*; IEEE: Pocono Manor, PA, USA, 2000; pp 334–338. https://doi.org/10.1109/SSAP.2000.870139.

(102) Georgiev, P.; Ralescu, A.; Ralescu, D. Cross-Cumulants Measure for Independence. *J. Stat. Plan. Inference* **2007**, *137* (3), 1085–1098. https://doi.org/10.1016/j.jspi.2006.06.034.

(103) Comon, P.; Jutten, C. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*; Academic press, 2010.

(104) Nuzillard, D.; Bourg, S.; Nuzillard, J.-M. Model-Free Analysis of Mixtures by NMR Using Blind Source Separation. *J. Magn. Reson.* **1998**, *133* (2), 358–363. https://doi.org/10.1006/jmre.1998.1481.

(105) Toumi, I.; Torrésani, B.; Caldarelli, S. Effective Processing of Pulse Field Gradient NMR of Mixtures by Blind Source Separation. *Anal. Chem.* **2013**, *85* (23), 11344–11351. https://doi.org/10.1021/ac402085x.

(106) Toumi, I.; Caldarelli, S.; Torrésani, B. A Review of Blind Source Separation in NMR Spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *81*, 37–64. https://doi.org/10.1016/j.pnmrs.2014.06.002.

(107) McCarty, R. J.; Ronghe, N.; Woo, M.; Alam, T. M. *Blind Source Separation for NMR Spectra with Negative Intensity*; preprint; CHEMISTRY, 2020. https://doi.org/10.20944/preprints202002.0113.v1.

(108) Robinette et al. - 2013 - Statistical Spectroscopic Tools for Biomarker Disc.Pdf.

(109) Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic [1] H NMR Data Sets. *Anal. Chem.* **2005**, *77* (5), 1282–1289. https://doi.org/10.1021/ac048630x.

(110) Ebrahimi, P.; Viereck, N.; Bro, R.; Engelsen, S. B. Chemometric Analysis of NMR Spectra. In *Modern Magnetic Resonance*; Webb, G. A., Ed.; Springer International Publishing: Cham, 2018; pp 1649–1668. https://doi.org/10.1007/978-3-319-28388-3_20.

(111) Huo, R.; Geurts, C.; Brands, J.; Wehrens, R.; Buydens, L. M. C. Real-Life Applications of the MULVADO Software Package for Processing DOSY NMR Data. *Magn. Reson. Chem.* **2006**, *44* (2), 110–117. https://doi.org/10.1002/mrc.1721.

(112) Salomonsen, T.; Jensen, H. M.; Larsen, F. H.; Steuernagel, S.; Engelsen, S. B. Direct Quantification of M/G Ratio from 13C CP-MAS NMR Spectra of Alginate Powders by Multivariate Curve Resolution. *Carbohydr. Res.* **2009**, *344* (15), 2014–2022. https://doi.org/10.1016/j.carres.2009.06.025.

(113) Nielsen, M. M.; Viereck, N.; Engelsen, S. B. PHYTIC ACID DEGRADATION BY PHYTASE – AS VIEWED BY 31P NMR AND MULTIVARIATE CURVE RESOLUTION. *Magn. Reson. Food Sci.* 9.

(114) Bruno, F.; Francischello, R.; Bellomo, G.; Gigli, L.; Flori, A.; Menichetti, L.; Tenori, L.; Luchinat, C.; Ravera, E. Multivariate Curve Resolution for 2D Solid-State NMR Spectra. *Anal. Chem.* **2020**, *92* (6), 4451–4458. https://doi.org/10.1021/acs.analchem.9b05420.

(115) Lee, D. D.; Seung, H. S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401* (6755), 788–791. https://doi.org/10.1038/44565.

(116) Sun, Y.; Babu, P.; Palomar, D. P. Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning. *IEEE Trans. Signal Process.* **2017**, *65* (3), 794–816. https://doi.org/10.1109/TSP.2016.2601299.

(117) Chapter 4: Majorization and Minorization. 29.

(118) Naderi, S.; He, K.; Aghajani, R.; Sclaroff, S.; Felzenszwalb, P. Generalized Majorization-Minimization. 10.

(119) Moreau, E.; Macchi, O. A One Stage Self-Adaptive Algorithm for Source Separation. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: Adelaide, SA, Australia, 1994; Vol. iii, p III/49-III/52. https://doi.org/10.1109/ICASSP.1994.390093.

(120) Févotte, C.; Idier, J. Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence. *ArXiv10101763 Cs* **2011**.

(121) Akoka, S.; Barantin, L.; Trierweiler, M. Concentration Measurement by Proton NMR Using the ERETIC Method. 4.

(122) Lee, Sueg-Geun. Heteronuclear Electronic Reference NMR Method for the Measurement of Concentration. *Bull. Korean Chem. Soc.* **2007**, *28* (10), 1635–1636. https://doi.org/10.5012/BKCS.2007.28.10.1635.

(123) Todeschini, R. *Introduzione alla Chemiometria.*, EdiSES.; 2003.

(124) Jacob, D.; Deborde, C.; Lefebvre, M.; Maucourt, M.; Moing, A. NMRProcFlow: A Graphical and Interactive Tool Dedicated to 1D Spectra Processing for NMR-Based Metabolomics. *Metabolomics* **2017**, *13* (4), 36. https://doi.org/10.1007/s11306-017-1178-y.

(125) Savorani, F.; Tomasi, G.; Engelsen, S. B. Icoshift: A Versatile Tool for the Rapid Alignment of 1D NMR Spectra. *J. Magn. Reson.* **2010**, *202* (2), 190–202. https://doi.org/10.1016/j.jmr.2009.11.012.

(126) Chaux, C.; Piersanti, E.; Shintu, L.; Yemloul, M. NMR Spectra (1H and HSQC) of Terpenes [Data Set]. In Faraday Discussions (v1.0, Vol. 218, Pp. 459–480). Zenodo. 5281/zenodo.4897068 2021.

(127) Huang, N. E.; Shen, Z.; Long, S. R.; Wu, M. C.; Shih, H. H.; Zheng, Q.; Yen, N.-C.; Tung, C. C.; Liu, H. H. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proc. R. Soc. Lond. Ser. Math. Phys. Eng. Sci.* **1998**, *454* (1971), 927. https://doi.org/10.1098/rspa.1998.0193.

(128) Bingol, K.; Li, D.-W.; Zhang, B.; Brüschweiler, R. Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. *Anal. Chem.* **2016**, *88* (24), 12411–12418. https://doi.org/10.1021/acs.analchem.6b03724.

(129) Crook, A. A.; Powers, R. Quantitative NMR-Based Biomedical Metabolomics: Current Status and Applications. *Molecules* **2020**, *25* (21), 5128. https://doi.org/10.3390/molecules25215128.

(130) Bingol, K.; Bruschweiler-Li, L.; Li, D.-W.; Brüschweiler, R. Customized Metabolomics Database for the Analysis of NMR ^{1}H–^{1}H TOCSY and ^{13}C–^{1}H HSQC-TOCSY Spectra of Complex Mixtures. *Anal. Chem.* **2014**, *86* (11), 5494–5501. https://doi.org/10.1021/ac500979g.

(131) Hansen, A. L.; Kupče, E.; Li, D.-W.; Bruschweiler-Li, L.; Wang, C.; Brüschweiler, R. 2D NMR-Based Metabolomics with HSQC/TOCSY NOAH Supersequences. *Anal. Chem.* **2021**, *93* (15), 6112–6119. https://doi.org/10.1021/acs.analchem.0c05205.

(132) Mahrous, E. A. Two Dimensional NMR Spectroscopic Approaches for Exploring Plant Metabolome: A Review. 13.

(133) Naanaa, W.; Nuzillard, J.-M. Blind Source Separation of Positive and Partially Correlated Data. *Signal Process.* **2005**, *85* (9), 1711–1722. https://doi.org/10.1016/j.sigpro.2005.03.006.

(134) Caldarelli, S. Chromatographic NMR: A Tool for the Analysis of Mixtures of Small Molecules. *Magn. Reson. Chem.* **2007**, *45* (S1), S48–S55. https://doi.org/10.1002/mrc.2143.

(135) Pages, G.; Delaurent, C.; Caldarelli, S. Simplified Analysis of Mixtures of Small Molecules by Chromatographic NMR Spectroscopy. *Angew. Chem. Int. Ed.* **2006**, *45* (36), 5950–5953. https://doi.org/10.1002/anie.200601622.

(136) Zhang, X.; Poniewierski, A.; Jelińska, A.; Zagożdżon, A.; Wisniewska, A.; Hou, S.; Hołyst, R. Determination of Equilibrium and Rate Constants for Complex Formation by Fluorescence Correlation Spectroscopy Supplemented by Dynamic Light Scattering and Taylor Dispersion Analysis. *Soft Matter* **2016**, *12* (39), 8186–8194. https://doi.org/10.1039/C6SM01791F.

(137) Hammouda, B. Temperature Effect on the Nanostructureof SDS Micelles in Water. *J. Res. Natl. Inst. Stand. Technol.* **2013**, *118*, 151. https://doi.org/10.6028/jres.118.008.

(138) Mirgorod, Y.; Chekadanov, A.; Dolenko, T. Structure of Micelles of Sodium Dodecyl Sulphate in Water: An X-Ray and Dynamic Light Scattering Study. *Chem. J. Mold.* **2019**, *14* (1), 107–119. https://doi.org/10.19261/cjm.2019.572.

(139) Clifford, J.; Pethica, B. A. The Self-Diffusion Coefficient of Sodium Dodecyl Sulfate Micelles. *J. Phys. Chem.* **1966**, *70* (10), 3345–3346. https://doi.org/10.1021/j100882a506.

(140) He, Z.; Alexandridis, P. Micellization Thermodynamics of Pluronic P123 (EO20PO70EO20) Amphiphilic Block Copolymer in Aqueous Ethylammonium Nitrate (EAN) Solutions. *Polymers* **2017**, *10* (1), 32. https://doi.org/10.3390/polym10010032.

(141) Lameiras, P.; Mougeolle, S.; Pedinielli, F.; Nuzillard, J.-M. Polar Mixture Analysis by NMR under Spin Diffusion Conditions in Viscous Sucrose Solution and Agarose Gel. *Faraday Discuss.* **2019**, *218*, 233–246. https://doi.org/10.1039/C8FD00226F.

(142) Lameiras, P.; Nuzillard, J.-M. Highly Viscous Binary Solvents: DMSO-D6/Glycerol and DMSO-D6/Glycerol-D8 for Polar and Apolar Mixture Analysis by NMR. *Anal. Chem.* **2016**, *88* (8), 4508–4515. https://doi.org/10.1021/acs.analchem.6b00481.

(143) Pedinielli, F.; Nuzillard, J.-M.; Lameiras, P. Mixture Analysis in Viscous Solvents by NMR Spin Diffusion Spectroscopy: ViscY. Application to High- and Low-Polarity Organic Compounds Dissolved in Sulfolane/Water and Sulfolane/DMSO-d_6 Blends. *Anal. Chem.* **2020**, *92* (7), 5191–5199. https://doi.org/10.1021/acs.analchem.9b05725.

(144) Jayaraman, V.; Ghosh, S.; Sengupta, A.; Srivastava, S.; Sonawat, H. M.; Narayan, P. K. Identification of Biochemical Differences between Different Forms of Male Infertility by Nuclear Magnetic Resonance (NMR) Spectroscopy. *J. Assist. Reprod. Genet.* **2014**, *31* (9), 1195–1204. https://doi.org/10.1007/s10815-014-0282-4.

(145) Fiehn, O. Metabolomics — the Link between Genotypes and Phenotypes. In *Functional Genomics*; Town, C., Ed.; Springer Netherlands: Dordrecht, 2002; pp 155–171. https://doi.org/10.1007/978-94-010-0448-0_11.

(146) Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic Profiling, Metabolomic and Metabonomic Procedures for NMR Spectroscopy of Urine, Plasma, Serum and Tissue Extracts. 12.

(147) Cox, D. G.; Oh, J.; Keasling, A.; Colson, K. L.; Hamann, M. T. The Utility of Metabolomics in Natural Product and Biomarker Characterization. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2014**, *1840* (12), 3460–3474. https://doi.org/10.1016/j.bbagen.2014.08.007.

(148) Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L. Systems Level Studies of Mammalian Metabolomes: The Roles of Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy. *Chem Soc Rev* **2011**, *40* (1), 387–426. https://doi.org/10.1039/B906712B.

(149) Metabolomics: Current Analytical Platforms and Methodologies. *TrAC Trends Anal. Chem.* **2005**, *24* (4), 285–294. https://doi.org/10.1016/j.trac.2004.11.021.

(150) Kim, H. K.; Choi, Y. H.; Verpoorte, R. NMR-Based Plant Metabolomics: Where Do We Stand, Where Do We Go? *Trends Biotechnol.* **2011**, *29* (6), 267–275. https://doi.org/10.1016/j.tibtech.2011.02.001.

(151) Lindon, J. C. Overview of NMR-Based Metabonomics. In *Encyclopedia of Spectroscopy and Spectrometry*; Elsevier, 2017; pp 517–526. https://doi.org/10.1016/B978-0-12-803224-4.00078-9.

(152) Meiboom, S.; Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Rev. Sci. Instrum.* **1958**, *29* (8), 688–691. https://doi.org/10.1063/1.1716296.

(153) Weininger, U.; Respondek, M.; Löw, C.; Akke, M. Slow Aromatic Ring Flips Detected Despite Near-Degenerate NMR Frequencies of the Exchanging Nuclei. *J. Phys. Chem. B* **2013**, *117* (31), 9241–9247. https://doi.org/10.1021/jp4058065.

(154) Mahi, L.; Duplan, J. C.; Fenet, B. Suppression of J-Modulation Effects for Resonance Editing by NMR Spin-Echo Experiments. *Chem. Phys. Lett.* **1993**, *211* (1), 27–30. https://doi.org/10.1016/0009-2614(93)80046-R.

(155) Garbow, J. R.; Weitekamp, D. P.; Pines, A. Bilinear Rotation Decoupling of Homonuclear Scalar Interactions. *Chem. Phys. Lett.* **1982**, *93* (5), 504–509. https://doi.org/10.1016/0009-2614(82)83229-6.

(156) Botana, A.; Aguilar, J. A.; Nilsson, M.; Morris, G. A. J-Modulation Effects in DOSY Experiments and Their Suppression: The Oneshot45 Experiment. *J. Magn. Reson.* **2011**, *208* (2), 270–278. https://doi.org/10.1016/j.jmr.2010.11.012.

(157) Barrère, C.; Thureau, P.; Thévand, A.; Viel, S. A Convenient Method for the Measurements of Transverse Relaxation Rates in Homonuclear Scalar Coupled Spin Systems. *Chem. Commun.* **2011**, *47* (32), 9209. https://doi.org/10.1039/c1cc13042k.

(158) Mandelshtam, V. A.; Van, Q. N.; Shaka, A. J. Obtaining Proton Chemical Shifts and Multiplets from Several 1D NMR Signals. *J. Am. Chem. Soc.* **1998**, *120* (46), 12161–12162. https://doi.org/10.1021/ja9824977.

(159) André, M.; Dumez, J.-N.; Rezig, L.; Shintu, L.; Piotto, M.; Caldarelli, S. Complete Protocol for Slow-Spinning High-Resolution Magic-Angle Spinning NMR Analysis of Fragile Tissues. *Anal. Chem.* **2014**, *86* (21), 10749–10754. https://doi.org/10.1021/ac502792u.

(160) van Zijl, P. C. M.; Moonen, C. T. W.; von Kienlin, M. Homonuclear J Refocusing in Echo Spectroscopy. *J. Magn. Reson. 1969* **1990**, *89* (1), 28–40. https://doi.org/10.1016/0022-2364(90)90159-7.

(161) Torres, A. M.; Zheng, G.; Price, W. S. J-Compensated PGSE: An Improved NMR Diffusion Experiment with Fewer Phase Distortions. *Magn. Reson. Chem.* **2009**, n/a-n/a. https://doi.org/10.1002/mrc.2555.

(162) Ouldamer, L.; Nadal-Desbarats, L.; Chevalier, S.; Body, G.; Goupille, C.; Bougnoux, P. NMR-Based Lipidomic Approach To Evaluate Controlled Dietary Intake of Lipids in Adipose Tissue of a Rat Mammary Tumor Model. *J. Proteome Res.* **2016**, *15* (3), 868–878. https://doi.org/10.1021/acs.jproteome.5b00788.

(163) Redfield, A. G. Nuclear Magnetic Resonance Saturation and Rotary Saturation in Solids. *Phys. Rev.* **1955**, *98* (6), 1787–1809. https://doi.org/10.1103/PhysRev.98.1787.

(164) Novoa-Carballal, R.; Fernandez-Megia, E.; Jimenez, C.; Riguera, R. NMR Methods for Unravelling the Spectra of Complex Mixtures. *Nat Prod Rep* **2011**, *28* (1), 78–98. https://doi.org/10.1039/C005320C.

(165) Shet, K.; Yoshihara, H.; Schooler, J.; Kurhanewicz, J.; Ries, M.; Li, X. T1-Rho Dispersion in Human OA Cartilage Specimens Using HRMAS Spectroscopy at 11.7T. 1.

(166) Tranchida, F.; Shintu, L.; Rakotoniaina, Z.; Tchiakpe, L.; Deyris, V.; Hiol, A.; Caldarelli, S. Metabolomic and Lipidomic Analysis of Serum Samples Following Curcuma Longa Extract Supplementation in High-Fructose and Saturated Fat Fed Rats. *PLOS ONE* **2015**, *10* (8), e0135948. https://doi.org/10.1371/journal.pone.0135948.

(167) Tranchida, F.; Rakotoniaina, Z.; Shintu, L.; Tchiakpe, L.; Deyris, V.; Yemloul, M.; Stocker, P.; Vidal, N.; Rimet, O.; Hiol, A.; Caldarelli, S. Hepatic Metabolic Effects of Curcuma Longa Extract Supplement in High-Fructose and Saturated Fat Fed Rats. *Sci. Rep.* **2017**, *7* (1), 5880. https://doi.org/10.1038/s41598-017-06220-0.

(168) Wishart, D. S. Current Progress in Computational Metabolomics. *Brief. Bioinform.* **2007**, *8* (5), 279–293. https://doi.org/10.1093/bib/bbm030.

(169) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. Assessment of PLSDA Cross Validation. *Metabolomics* **2008**, *4* (1), 81–89. https://doi.org/10.1007/s11306-007-0099-6.

(170) Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. 16.

(171) Ban, D.; Gossert, A. D.; Giller, K.; Becker, S.; Griesinger, C.; Lee, D. Exceeding the Limit of Dynamics Studies on Biomolecules Using High Spin-Lock Field Strengths with a Cryogenically Cooled Probehead. *J. Magn. Reson.* **2012**, *221*, 1–4. https://doi.org/10.1016/j.jmr.2012.05.005.

(172) Holly, R.; Peemoeller, H.; Choi, C.; Pintar, M. M. Proton Rotating Frame Spin-Lattice Relaxation Study of Slow Motion of Pore Water. *J. Chem. Phys.* **1998**, *108* (10), 4183–4188. https://doi.org/10.1063/1.475816.

(173) Haukaas, T. H.; Moestue, S. A.; Vettukattil, R.; Sitter, B.; Lamichhane, S.; Segura, R.; Giskeødegård, G. F.; Bathen, T. F. Impact of Freezing Delay Time on Tissue Samples for Metabolomic Studies. *Front. Oncol.* **2016**, *6*. https://doi.org/10.3389/fonc.2016.00017.

(174) Lewandowski, J. R.; Sass, H. J.; Grzesiek, S.; Blackledge, M.; Emsley, L. Site-Specific Measurement of Slow Motions in Proteins. *J. Am. Chem. Soc.* **2011**, *133* (42), 16762–16765. https://doi.org/10.1021/ja206815h.

(175) Ma, P.; Haller, J. D.; Zajakala, J.; Macek, P.; Sivertsen, A. C.; Willbold, D.; Boisbouvier, J.; Schanda, P. Probing Transient Conformational States of Proteins by Solid-State R $_{1\rho}$ Relaxation-Dispersion NMR Spectroscopy. *Angew. Chem. Int. Ed.* **2014**, *53* (17), 4312–4317. https://doi.org/10.1002/anie.201311275.

(176) Tsutsumi, A. Side Chain Dynamics in Poly(y-Benzyl L-Glutamate) as Studied by High-Resolution Solid State 13C Nuclear Magnetic Relaxation in Rotating Frame. *Polym J* **1993**, *25* (2), 9.

(177) Krushelnitsky, A.; Zinkevich, T.; Reif, B.; Saalwächter, K. Slow Motions in Microcrystalline Proteins as Observed by MAS-Dependent 15N Rotating-Frame NMR Relaxation. *J. Magn. Reson.* **2014**, *248*, 8–12. https://doi.org/10.1016/j.jmr.2014.09.007.

(178) Kurbanov, R.; Zinkevich, T.; Krushelnitsky, A. The Nuclear Magnetic Resonance Relaxation Data Analysis in Solids: General R_1 / $R_{1\rho}$ Equations and the Model-Free Approach. *J. Chem. Phys.* **2011**, *135* (18), 184104. https://doi.org/10.1063/1.3658383.

(179) Steiner, E.; Yemloul, M.; Guendouz, L.; Leclerc, S.; Robert, A.; Canet, D. NMR Relaxometry: Spin Lattice Relaxation Times in the Laboratory Frame versus Spin Lattice Relaxation Times in the Rotating Frame. *Chem. Phys. Lett.* **2010**, *495* (4–6), 287–291. https://doi.org/10.1016/j.cplett.2010.06.064.

(180) Gan, Z.-H.; Grant, D. M. Rotational Resonance in a Spin-Lock Field for Solid State NMR. *Chem. Phys. Lett.* **1990**, *168* (3–4), 304–308. https://doi.org/10.1016/0009-2614(90)85615-J.

(181) Gan, Z.; Grant, D. M.; Ernst, R. R. NMR Chemical Shift Anisotropy Measurements by RF Driven Rotary Resonance. *Chem. Phys. Lett.* **1996**, *254* (5–6), 349–357. https://doi.org/10.1016/0009-2614(96)00268-0.

(182) Stark, D. A.; Timar, B.; Madar, Z. Adaptation of Sprague Dawley Rats to Long-Term Feeding of High Fat or High Fructose Diets. *Eur. J. Nutr.* **2000**, *39* (5), 6.

(183) Gayles, E. C.; Pagliassotti, M. J.; Prach, P. A.; Koppenhafer, T. A.; Hill, J. O. Contribution of Energy Intake and Tissue Enzymatic Profile to Body Weight Gain in High-Fat-Fed Rats. *Am. J. Physiol.-Regul. Integr. Comp. Physiol.* **1997**, *272* (1), R188–R194. https://doi.org/10.1152/ajpregu.1997.272.1.R188.

(184) Shekar, S. C.; Lee, D.-K.; Ramamoorthy, A. Chemical Shift Anisotropy and Offset Effects in Cross Polarization Solid-State NMR Spectroscopy. *J. Magn. Reson.* **2002**, *157* (2), 223–234. https://doi.org/10.1006/jmre.2002.2587.

(185) Piersanti, E.; Rezig, L.; Tranchida, F.; El-Houri, W.; Abagana, S. M.; Campredon, M.; Shintu, L.; Yemloul, M. Evaluation of the Rotating-Frame Relaxation ($T_{1\rho}$) Filter and Its Application in Metabolomics as an Alternative to the Transverse Relaxation (T_2) Filter. *Anal. Chem.* **2021**, *93* (25), 8746–8753. https://doi.org/10.1021/acs.analchem.0c05251.

(186) Eicher, J. J. Determining Enzyme Kinetics for Systems Biology with Nuclear Magnetic Resonance Spectroscopy. **2012**, 26.

(187) Blass, D. A.; Adams, E. Polarimetry as a General Method for Enzyme Assays. 10.

(188) Ma, J.; McLeod, S.; MacCormack, K.; Sriram, S.; Gao, N.; Breeze, A. L.; Hu, J. Real-Time Monitoring of New Delhi Metallo-β-Lactamase Activity in Living Bacterial Cells by [1] H NMR Spectroscopy. *Angew. Chem. Int. Ed.* **2014**, *53* (8), 2130–2133. https://doi.org/10.1002/anie.201308636.

(189) Werner, R. M.; Johnson, A. [31] P NMR of the Pyruvate Kinase Reaction: An Undergraduate Experiment in Enzyme Kinetics: [31] P NMR of the Pyruvate Kinase Reaction. *Biochem. Mol. Biol. Educ.* **2017**, *45* (6), 509–514. https://doi.org/10.1002/bmb.21079.

(190) Exnowitz, F.; Meyer, B.; Hackl, T. NMR for Direct Determination of Km and Vmax of Enzyme Reactions Based on the Lambert W Function-Analysis of Progress Curves. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **2012**, *1824* (3), 443–449. https://doi.org/10.1016/j.bbapap.2011.10.011.

(191) Kehlbeck, J. D.; Slack, C. C.; Turnbull, M. T.; Kohler, S. J. Exploring the Hydrolysis of Sucrose by Invertase Using Nuclear Magnetic Resonance Spectroscopy: A Flexible Package of Kinetic Experiments. *J Chem Educ* **2014**, 5.

(192) Olsen2010.Pdf.

(193) Jiang, Y.; McKinnon, T.; Varatharajan, J.; Glushka, J.; Prestegard, J. H.; Sornborger, A. T.; Schüttler, H.-B.; Bar-Peled, M. Time-Resolved NMR: Extracting the Topology of Complex Enzyme Networks. *Biophys. J.* **2010**, *99* (7), 2318–2326. https://doi.org/10.1016/j.bpj.2010.08.014.

(194) Stejskal, E. O.; Tanner, J. E. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *J. Chem. Phys.* **1965**, *42* (1), 288–292. https://doi.org/10.1063/1.1695690.

(195) Antalek, B. Using PGSE NMR for Chemical Mixture Analysis: Quantitative Aspects. *Concepts Magn. Reson. Part A* 17.

(196) Viel, S.; Caldarelli, S. Improved 3D DOSY-TOCSY Experiment for Mixture Analysis. *Chem. Commun.* **2008**, No. 17, 2013. https://doi.org/10.1039/b802789g.

(197) Nilsson, M.; Gil, A. M.; Delgadillo, I.; Morris, G. A. Improving Pulse Sequences for 3D Diffusion-Ordered NMR Spectroscopy: 2DJ-IDOSY. 5.

(198) Nilsson, M.; Morris, G. A. Improving Pulse Sequences for 3D DOSY: Convection Compensation. *J. Magn. Reson.* **2005**, *177* (2), 203–211. https://doi.org/10.1016/j.jmr.2005.07.019.

(199) Wu, D.; Chen, A.; Johnson, C. S. Three-Dimensional Diffusion-Ordered NMR Spectroscopy: The Homonuclear COSY-DOSY Experiment. 4.

(200) Jerschow, A.; Müller, N. 3D Diffusion-Ordered TOCSY for Slowly Diffusing Molecules. *J. Magn. Reson. A* **1996**, *123* (2), 222–225. https://doi.org/10.1006/jmra.1996.0241.

(201) G. N., M. R.; Caldarelli, S. Demixing of Severely Overlapping NMR Spectra through Multiple-Quantum NMR. *Anal. Chem.* **2010**, *82* (8), 3266–3269. https://doi.org/10.1021/ac100009y.

(202) Manjunatha Reddy, G. N.; Caldarelli, S. Improved Excitation Uniformity in Multiple-Quantum NMR Experiments of Mixtures: Uniform MaxQ. *Magn. Reson. Chem.* **2013**, *51* (4), 240–244. https://doi.org/10.1002/mrc.3938.

(203) Manjunatha Reddy, G. N.; Yemloul, M.; Caldarelli, S. Combined Maximum-Quantum and DOSY 3D Experiments Provide Enhanced Resolution for Small Molecules in Mixtures: MaxQ-MaxQDOSY. *Magn. Reson. Chem.* **2017**, *55* (5), 492–497. https://doi.org/10.1002/mrc.4465.

(204) Reddy, G. N. M.; Caldarelli, S. Identification and Quantification of EPA 16 Priority Polycyclic Aromatic Hydrocarbon Pollutants by Maximum-Quantum NMR. *The Analyst* **2012**, *137* (3), 741–746. https://doi.org/10.1039/C1AN16047H.

(205) Wokaun, A.; Ernst, R. R. The Use of Multiple Quantum Transitions for Relaxation Studies in Coupled Spin Systems. *MOLECULAR PHYSICS*. 36th ed. 1978, pp 317–341.

(206) Reddy, G. N. M.; Caldarelli, S. Identification and Quantification of EPA 16 Priority Polycyclic Aromatic Hydrocarbon Pollutants by Maximum-Quantum NMR. *The Analyst* **2012**, *137* (3), 741–746. https://doi.org/10.1039/C1AN16047H.

(207) Urresti, S.; Lafond, M.; Johnston, E. M.; Derikvand, F.; Ciano, L.; Berrin, J.-G.; Henrissat, B.; Walton, P. H.; Davies, G. J.; Brumer, H. Structure–Function Characterization Reveals New Catalytic Diversity in the Galactose Oxidase and Glyoxal Oxidase Family. *Nat. Commun.* 13.

(208) A. Bax; De Jong, P. G.; Mehlkopf, A. F.; Smidt, J. SEPARATION OF THE DIFFERENT ORDERS OF NMR MULTIPLEX-QUANTUM TRANSITIONS BY THE USE OF PULSED FIELD GRADIENTS. *CHEMICAL PHYSICS LETTERS*. February 1, 1980, pp 567–570.

(209) Reddy, G. N. M.; Caldarelli, S. Improved Excitation Uniformity in Multiple-quantum NMR Experiments of Mixtures. *Magn Reson Chem* **2013**, 5.

(210) Manjunatha Reddy, G. N.; Caldarelli, S. Improved Excitation Uniformity in Multiple-Quantum NMR Experiments of Mixtures: Uniform MaxQ. *Magn. Reson. Chem.* **2013**, *51* (4), 240–244. https://doi.org/10.1002/mrc.3938.

(211) Le Guennec, A.; Dumez, J.-N.; Giraudeau, P.; Caldarelli, S. Resolution-Enhanced 2D NMR of Complex Mixtures by Non-Uniform Sampling: Resolution-Enhanced 2D NMR of Complex Mixtures by Non-Uniform Sampling. *Magn. Reson. Chem.* **2015**, *53* (11), 913–920. https://doi.org/10.1002/mrc.4258.

(212) Nilsson, M. The DOSY Toolbox: A New Tool for Processing PFG NMR Diffusion Data. *J. Magn. Reson.* **2009**, *200* (2), 296–302. https://doi.org/10.1016/j.jmr.2009.07.022.

(213) Gozansky, E. K.; Gorenstein, D. G. DOSY-NOESY: Diffusion-Ordered NOESY. *J. Magn. Reson. B* **1996**, *111* (1), 94–96. https://doi.org/10.1006/jmrb.1996.0066.

(214) Zax, D.; Pines, A. Study of Anisotropic Diffusion of Oriented Molecules by Multiple Quantum Spin Echoes. *J. Chem. Phys.* **1983**, *78* (10), 6333–6334. https://doi.org/10.1063/1.444559.

(215) Martin, J. F.; Selwyn, L. S.; Vold, R. R.; Vold, R. L. The Determination of Translational Diffusion Constants in Liquid Crystals from Pulsed Field Gradient Double Quantum Spin Echo Decays. *J. Chem. Phys.* **1982**, *76* (5), 2632–2634. https://doi.org/10.1063/1.443243.

(216) Kay, L. E.; Prestegard, J. H. Spin-Lattice Relaxation Rates of Coupled Spins from 2D Accordion Spectroscopy. *J. Magn. Reson. 1969* **1988**, *77* (3), 599–605. https://doi.org/10.1016/0022-2364(88)90021-2.

(217) van Dam, L.; Andreasson, B.; Nordenskiöld, L. Multiple-Quantum Pulsed Gradient NMR Diffusion Experiments on Quadrupolar (I >) Spins. *Chem. Phys. Lett.* **1996**, *262* (6), 737–743. https://doi.org/10.1016/S0009-2614(96)01160-8.

(218) Sotak, C. H. Multiple Quantum NMR Spectroscopy Methods for Measuring the Apparent Self-Diffusion Coefficient OfIn Vivo Lactic Acid. *NMR Biomed.* **1991**, *4* (2), 70–72. https://doi.org/10.1002/nbm.1940040207.

(219) Hedin, N.; Yu, T. Y.; Furó, I. Growth of $C_{12}E_8$ Micelles with Increasing Temperature. A Convection-Compensated PGSE NMR Study. *Langmuir* **2000**, *16* (19), 7548–7550. https://doi.org/10.1021/la000595b.

(220) Jerschow, A.; Müller, N. Convection Compensation in Gradient Enhanced Nuclear Magnetic Resonance Spectroscopy. *J. Magn. Reson.* **1998**, *132* (1), 13–18. https://doi.org/10.1006/jmre.1998.1400.

(221) Momot, K. I.; Kuchel, P. W. Convection-Compensating Diffusion Experiments with Phase-Sensitive Double-Quantum Filtering. *J. Magn. Reson.* **2005**, *174* (2), 229–236. https://doi.org/10.1016/j.jmr.2005.02.003.

(222) Zhang, X.; Li, C.-G.; Ye, C.-H.; Liu, M.-L. Determination of Molecular Self-Diffusion Coefficient Using Multiple Spin-Echo NMR Spectroscopy with Removal of Convection and Background Gradient Artifacts. *Anal. Chem.* **2001**, *73* (15), 3528–3534. https://doi.org/10.1021/ac0101104.

(223) Ribeaucourt, D.; Höfler, G. T.; Yemloul, M.; Bissaro, B.; Lambert, F.; Berrin, J.-G.; Lafond, M.; Paul, C. E. Tunable Production of (*R*)- or (*S*)-Citronellal from Geraniol via a Bienzymatic Cascade Using a Copper Radical Alcohol Oxidase and Old Yellow Enzyme. *ACS Catal.* **2022**, 1111–1116. https://doi.org/10.1021/acscatal.1c05334.

(224) Yan, D.; Wu, T.; Liu, Y.; Gao, Y. An Efficient Sparse-Dense Matrix Multiplication on a Multicore System. In *2017 IEEE 17th International Conference on Communication Technology (ICCT)*; IEEE: Chengdu, 2017; pp 1880–1883. https://doi.org/10.1109/ICCT.2017.8359956.

(225) Penfield, P. Information and Entropy Lecture Notes; pp 104–118.

(226) Lee, D. D.; Seung, H. S. Algorithms for Non-Negative Matrix Factorization. 7.

(227) Chouzenoux, E.; Pesquet, J.-C.; Repetti, A. A Block Coordinate Variable Metric Forward–Backward Algorithm. *J. Glob. Optim.* **2016**, *66* (3), 457–485. https://doi.org/10.1007/s10898-016-0405-9.

(228) Searcóid, M. Ó. Lipschitz Functions, Metric Spaces. Springer-Verlag 2006.

(229) Cobas, J. C.; Tahoces, P. G.; Martin-Pastor, M.; Penedo, M.; Javier Sardina, F. Wavelet-Based Ultra-High Compression of Multidimensional NMR Data Sets. *J. Magn. Reson.* **2004**, *168* (2), 288–295. https://doi.org/10.1016/j.jmr.2004.03.016.

(230) Shao, X.; Gu, H.; Wu, J.; Shi, Y. Resolution of the NMR Spectrum Using Wavelet Transform. *Appl. Spectrosc.* **2000**, *54* (5), 731–738. https://doi.org/10.1366/0003702001949997.

(231) Kopriva, I.; Jerić, I.; Smrečki, V. Extraction of Multiple Pure Component 1H and 13C NMR Spectra from Two Mixtures: Novel Solution Obtained by Sparse Component Analysis-Based Blind Decomposition. *Anal. Chim. Acta* **2009**, *653* (2), 143–153. https://doi.org/10.1016/j.aca.2009.09.019.

(232) Cherni, A.; Piersanti, E.; Anthoine, S.; Chaux, C.; Shintu, L.; Yemloul, M.; Torrésani, B. Challenges in the Decomposition of 2D NMR Spectra of Mixtures of Small Molecules. *Faraday Discuss.* **2019**, *218*, 459–480. https://doi.org/10.1039/C9FD00014C.

Table of Figures

Table of Tables

Acknowledgements

It is my duty to dedicate this space of my manuscript to the people who have contributed, with their support and all the discussions and advice, to its realization, but also to all the people met during this professional experience, allowing me to be the person I am today.

*Firstly, I would like to thank the members of the **jury**, **Prof. Jean-Marc Nuzillard, Prof. Bruno Torrésani**, **Dr. Laurent Duval**, for their generous involvement, taking their time to judge my work. More particularly, **Dr. Jonathan Farjon** and **Prof. Sabine Bouguet-Bonnet**, for being the reviewers of this manuscript. I thank them in advance for the corrections and advice they will offer me allowing me to improve.*

*I would like to express my sincere gratitude to **Dr. Mehdi Yemloul**, with his enthusiasm, patience, dedication and human qualities, for helping and teaching me during this Ph.D. work, sometimes with difficulties, always encouraging me to continue on my path.*

*I am grateful to our BIFROST project leader, **Dr. Caroline Chaux**, who supervised me for the mathematical part, with her motivation and investment in the work, constantly followed me with coding and mathematics, two different fields for me.*

*If all this was possible it is thanks to **Prof. Stefano Caldarelli,** my previous thesis director, who recruited me, giving me the opportunity to become part of his research team.*

*I would like to thank **Prof. Mylène Campredon** to have accepted to direct afterwards my PhD thesis.*

*If this work was possible, indeed, is thank to **I2M team**, with **Prof. Bruno Torrésani**, in particular with all his teachings, **Dr. Sandrine Anthoine, Prof. Frédéric Richard** and **Dr.**

*A non-negligible support came from two physicists, my best Erasmus friends, **Martina Zanetti** and **Giuseppe Torrisi**, always there to help me in the math part.*

*Thanks to my **Italian and Parisian friends** of always for the moments of escape from the thesis, and because even with just a message, a phone call, a small gesture they were close to me.*

*Finally, I must express my very profound gratitude to my **parents, Gianni** and **Anna,** and **my brother Leonardo,** for providing me with unfailing support and continuous encouragement throughout my years of studies, to make my dreams come true (despite the distance and Covid-19). This accomplishment would not have been possible without them. Thank you.*

Finally, I thank myself for my tenacity and my stubbornness, my ability to adapt and the strong side of my character, to always lift myself up with a smile even after difficult moments, and to have completed what I had set for myself, despite the sacrifices, despite the fact that on many occasions the simplest thing was to give up everything.

Elena

Résumé

Cette thèse vise au développement de méthodes de traitement du signal pour l'analyse de mélanges de petites molécules. La première partie est centrée sur l'évaluation d'algorithmes de séparation aveugle de sources (BSS) sur des données RMN 1D et 2D. Les résultats sur les données simulées fournissent une bonne estimation des empreintes spectrales et des concentrations des composants dans les mélanges. Pour les données réelles, le prétraitement est important comme l'alignement des spectres. Le modèle de mélange instantané linéaire s'est avéré incomplet pour décrire les ensembles de données. Pour une évaluation quantitative des performances, nous avons proposé des paramètres utilisés pour le BSS dans d'autres domaines que la spectroscopie : l'indice d'Amari, le SIR et le SDR. Leur pertinence pour ce type de données doit être validée avec un ensemble de données plus large.

Ensuite, deux méthodes de RMN sont évaluées dans deux applications originales : la première concerne la mise en œuvre et l'évaluation du filtre de relaxation $T_{1\rho}$ en métabolomique pour supprimer les signaux des macromolécules. Ce filtre est bien adapté aux échantillons étudiés en métabolomique, qu'ils soient liquides ou tissulaires, avec des résultats similaires à ceux obtenus avec le filtre de relaxation T_2. La seconde concerne l'application de l'approche Maximum Quantum (MaxQ-NMR) pour simplifier les spectres de mélanges issus de réactions enzymatiques. La présence d'enzymes et de complexes paramagnétiques conduit à une réduction significative des temps de relaxation nécessitant un réglage fin des paramètres expérimentaux pour trouver un compromis entre excitation uniforme des cohérences et perte de signal.

<u>Mots clés</u> : analyse de mélanges, traitement du signal, Résonance Magnétique Nucléaire, métabolomique.

Abstract

This thesis aims to develop signal processing methods for the analysis of mixtures of small molecules. The first part is focused on the evaluation of blind source separation (BSS) algorithms on 1D and 2D NMR data. The results on simulated data provide a good estimate of the spectral fingerprints and the concentrations of the components in the mixtures. For the real data, pre-processing steps are very important such as the alignment of the spectra. The linear instantaneous mixing model was found to be incomplete to accurately describe the datasets. For a quantitative evaluation of performances, we have proposed parameters used for the BSS on other fields than spectroscopy: the Amari index, SIR and SDR. Their relevance to this type of data has to be validated with a larger dataset.

In the second part, two NMR methods are evaluated in two original applications: the first concerns the implementation and evaluation of the $T_{1\rho}$ relaxation filter in metabolomics to suppress large signals from macromolecules. It is shown here that this filter is well suited for samples studied in metabolomics, whether liquids or tissues, with results similar to those obtained with the T_2 relaxation filter. The second concerns the application of the Maximum Quantum (MaxQ-NMR) approach to simplify the NMR spectra of mixtures resulting from enzymatic reactions. The presence of enzymes and paramagnetic complexes leads to a significant reduction in relaxation times requiring fine adjustment of experimental parameters to find a compromise between uniform excitation of coherences and loss of signal.

<u>Keywords</u>: mixture analysis, signal processing, Nuclear Magnetic Resonance, metabolomics.