Measurement Informatics Approaches for NMR Signal Deconvolution and Data-Driven Analysis Toward Molecular Complexity

（分子複雑系の核磁気共鳴信号分離およびデータ駆動型分析のための

計測インフォマティクスアプローチ）

Laboratory of Metabolic Balance of Ecosystem
Division of Biofunctional Systems
Department of Bioengineering Sciences
Graduate School of Bioagricultural Sciences
Nagoya University

名古屋大学　大学院生命農学研究科　生命技術科学専攻

生命機能システム学講座　環境調和システム研究室

Shunji Yamada

山田　隼嗣

March 2021

2021 年 3 月

Abstract

In order to solve the environmental, resource and health problems of water, food, materials, and energy, which are important issues in modern society, there is a demand for measurement technology for molecular complex systems. Therefore, mixtures composed of various molecules and higher-order structures of materials development of measurement technology is required. Nuclear magnetic resonance (NMR) spectroscopy characterizes molecular complex systems by using various acquisition parameters and pulse sequences to provide useful data on the chemical structure and molecular motility of non-invasive samples at atomic resolution. With the development of information technology in recent years, measurement informatics approaches for effectively utilizing accumulated NMR data are becoming more and more important. However, NMR spectrum analysis of the mixture is difficult due to noise and signal overlapping, and requires a lot of labor. Therefore, there is a need for research on data quality control, noise reduction, data cleansing such as signal separation, which is the previous stage of spectrum analysis, and data-driven analysis that makes use of the data accumulated over many years.

In this study, I first examined a signal assignment method that used a pre-processing method to emphasize and separate peaks for a broad NMR spectrum. I also performed noise factor analysis to investigate the relationship between the measurement parameters of NMR data accumulated in the laboratory and other institutions and the signal-to-noise ratio (SNR). Therefore, I have developed a signal deconvolution method that combines short-time Fourier transform (STFT) and probabilistic sparse matrix factorization (PSMF) for solution NMR signals containing different motility (T_2^* relaxation, which is referred to decay of transverse magnetization caused by a combination of transverse relaxation and magnetic field inhomogeneity) components and noise. Furthermore, based on this theory, the application of the signal deconvolution method was examined for solid-state NMR (ssNMR) signals of multi-component materials having multiple domain structures or components in the solid-state. Subsequently, the application of Generative Topographic Mapping Regression (GTMR) was examined in order to predict NMR signals and physical properties as descriptors of components and structures. Based on the above method, I have developed a measurement informatics approach useful for NMR analysis of molecular complex systems.

Firstly, the pre-processing method and signal assignment method for low- and high-field NMR analysis of molecular complex systems is developed. In NMR analysis

of molecular complex systems, various sample preparation methods and pulse sequences are used due to the variety of physicochemical properties of molecules. However, NMR signal assignment of a mixture is difficult due to signal overlapping problems and lack of reference spectra and signal assignment tools. Therefore, in this study, in order to support signal assignment in low-resolution NMR spectra analysis of molecular complex systems from small molecules to macromolecules and lipids, InterSpin is developed a web tool consisting of wide spectrum pre-processing, signal assignment, and database (DB). I have developed a combined method of SENSI, which improves sensitivity by spectral integration, and PKSP, which separates signals, as pre-processing tools that support the analysis of broad spectra obtained from low-field tabletop NMR and ssNMR. PKSP has implemented non-negative sparse coding (NNSC) as a new NMR signal separation method, enabling faster and more accurate signal separation than conventional methods such as NMF. Furthermore, by combining the coefficient of variation (CV) of each peak obtained by SENSI and the separation signal obtained by PKSP, the broad spectrum obtained by low-field NMR of fish dishes and ssNMR of *Euglena* is assigned. Extensive research on food, materials, environment, health, etc. requires diverse standard spectra. However, DB or signal assignment method has not been established for solution NMR in solid ^{13}C CP-MAS or DMSO-d6 solvents with similar structures. Therefore, a new SpinLIMS DB including various sample spectra (^{1}H-^{13}C correlation, ^{1}H-*J* resolved, ^{13}C CP-MAS) from small molecules to macromolecules and lipids in solid and solution states (D_2O, MeOD-d4, DMSO-d6 as solvents) is developed. This DB was constructed to enable signal assignment in InterSpin for various samples. Based on this DB, I developed SpinMacro, which assigns signals of macromolecules and lipids, and automated the signal assignment of solid ^{13}C peaks and ^{1}H-^{13}C correlation peaks in DMSO-d6 solvents. Furthermore, I have developed InterAnalysis that narrows down candidate molecules by integrating the ^{1}H-^{13}C correlation peaks and the ^{1}H-*J* resolved peaks, and streamlined signal assignment. The pre-processing method and signal assignment method for NMR analysis of molecular complex systems have solved the problem of signal overlapping in low-field NMR and ssNMR, and have advanced NMR signal assignment using DB.

Secondly, NMR signal deconvolution method and noise factor analysis method combining STFT and PSMF is proposed. In data-driven analysis, data quality is important because it affects results. However, in the field of NMR, data cleansing methods such as quality control, noise reduction, and signal separation of accumulated NMR data have not been established. Therefore, in this study, I focused on the measurement parameters and noise of NMR data, and developed a data cleansing method by free induction decay (FID) signal separation and noise factor analysis of one-dimensional NMR. In order to reduce

noise and separate signals, I examined the use of spectral changes at each time by dividing the FID at regular intervals by applying STFT. Based on the characteristics of signal intensity attenuation associated with T_2^* relaxation on the time axis added by STFT, matrix factorization was able to distinguish between signals and noise at individual frequencies. As a matrix factorization method, PSMF was able to separate signals and noise better than other methods such as NMF. The new signal separation method that combines STFT and PSMF, unlike the conventional noise reduction method that uses only multivariate analysis, does not require the FID of many samples or measurement parameters, and enables noise reduction and signal separation. By this method, the noise of the FID signal was separated and the SNR of the spectrum was improved about 3 times. The diffusion-edited NMR spectrum could be separated into signals of macromolecules, lipids and small molecules due to the difference in T_2^* of each frequency component. In order to utilize the accumulated NMR data, it is necessary to confirm the data quality. Therefore, as a result of noise factor analysis to investigate the relationship between the measurement parameters of NMR data and SNR, the number of scans was the main factor that reduced SNR when solvent suppression was insufficient. Signal deconvolution and noise factor analysis of solution NMR using STFT solved the problem of signal overlapping and advanced data-driven analysis of solution NMR.

Finally, the signal deconvolution method and signal/physical property prediction method for ssNMR of multi-component materials is demonstrated. In recent years, due to global issues such as marine pollution of marine plastics, waste treatment, and global warming, research into a low-carbon society has been emphasized. Microbial products and plant biomass as alternatives to petroleum resources can be used in the production of materials such as plastics and raw materials. Polymers such as polylactic acid (PLA), poly-ε-caprolactone (PCL), and cellulose are molecular complex systems with multiple domains or components and are used as materials with various properties. Microorganisms and plant biomass need to be analyzed as a biochemical system composed of multiple components, including lipids and macromolecules with multiple domains. Therefore, it is necessary to develop an ssNMR analysis approach for multi-component materials such as microbial products, plant biomass and plastics. Therefore, in this study, I developed a signal deconvolution method for ssNMR. I investigated a novel signal deconvolution method using STFT and non-negative tensors and matrix factorization (NTF, NMF). By this method, the ^{13}C CP-MAS spectrum in the cellulose decomposition process could be separated into cellulose, proteins, and lipids signals by the difference in T_2^* by STFT and NTF. In addition, the anisotropy spectrum of PCL could be separated into crystalline and amorphous signals by STFT and NMF. As an alternative

to decoupling, which is applied to remove anisotropy during measurement, signal deconvolution of anisotropy measurement data by computational scientific methods has been made possible. I also examined GTMR as a new method for visualizing and predicting NMR signals and physical properties using STFT data. GTMR was able to predict the NMR signal intensity of acetic acid and CO_2 as the products in the cellulose degradation process. In addition, by applying GTMR to the NMR signal of plastics and the integrated data of each physical property, it was possible to predict the NMR signal with the desired physical properties (glass transition point, melting point, degradation temperature). The ssNMR signal deconvolution and prediction method using STFT solved the signal overlapping problem and enabled the characterization and design of multi-component materials.

By each of the above methods, for NMR analysis of molecular complexity, the signal deconvolution and data-driven analysis related to signal assignment and prediction of higher-order structure and physical properties has been proposed. The measurement informatics approach developed in this research is expected to contribute to data-driven research, development, production, quality control, etc. of molecular complex systems in various fields such as health, food, materials, and the environment.

Contents

Part I

General Introduction

Chapter 1

Introduction

1.1 Concepts of Measurements Informatics in NMR Toward Molecular Complexity

The development of analytical approaches for understanding complex systems in natural science is a pivotal challenge for finding innovative solutions in modern society. In chemistry, a complex system is one whose evolution is very sensitive to initial conditions or to small perturbations, one in which the number of independent interacting components is large, or one in which there are multiple pathways by which the system can evolve[1]. In this thesis work, “molecular complexity” means complex structure and various composition of natural mixtures such as crude biological extracts, geochemical samples, and intact cells and tissues as well as materials with multi-domain and component[2].

Until 1960, Analysis of mixtures need physical separation such as filtration, distillation, fractionation, recrystallization, extraction, sublimation, chromatography, relied on elemental analysis[3]. Those approaches were not able to evaluate natural structure and components of intact samples. The advent of measurement techniques such as ultraviolet (UV), infrared (IR), mass spectrometry (MS) and nuclear magnetic resonance (NMR) changed mixture analysis. Among these methods, NMR is the most powerful non-destructive measurements open a window into exploring intact complex mixtures by combining some unique data processing methods.

Successful examples of the collaboration between natural science and informatics are bioinformatics and materials informatics[4]. From a different point of view, these successes were derived as a result of the combination of informatics and unique measurement techniques in each field of study. From this viewpoint, “measurement informatics”, which focuses on measurement and data utilization and is widely targeted in natural science such as food, life, environment, and materials, is a new research field.

This field will be formed by the fusion of analytical chemistry and informatics. This approach attempts to effectively utilize measurement data by circulating a series of data starting from the measurement of samples, rather than simply measurement and analysis (Figure 1). This field includes research on data management, pre-processing of data analysis, evaluation, prediction, and control of structure, composition, and condition in the molecule and natural phenomenon.

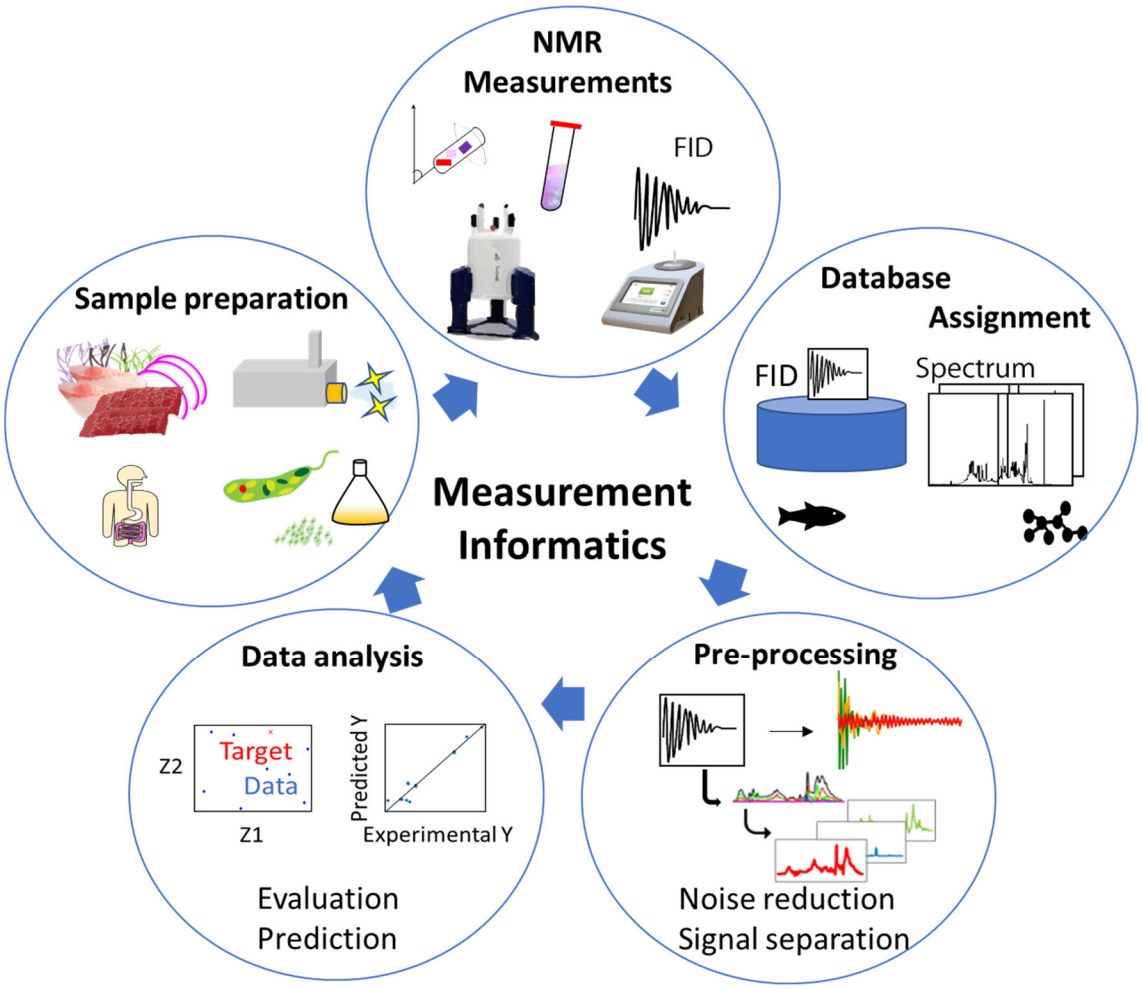


Figure 1. Concept diagram of measurement informatics in NMR toward molecular complexity. This figure shows the concept of measurement informatics, which consists of a cycle of sample preparation, solid-state, solution and low-field NMR measurements, database construction, data preprocessing, and data-driven analysis. These approaches are repeated based on the evaluation and prediction.

1.2 Challenges of Measurements Informatics Approaches in NMR Toward Molecular Complexity

In order to solve the environmental, resource and health problems of water, food, materials, and energy, which are important issues in modern society, I targeted molecular complex systems such as mixtures composed of various molecules and higher-order structures of materials. Development of measurement technology is required. Nuclear magnetic resonance (NMR) spectroscopy characterizes molecular complex systems by using various acquisition parameters and pulse sequences to provide useful data on the chemical structure and molecular motility of non-invasive samples at atomic resolution. NMR offers to deal with various and diverse samples from polar to non-polar solvent systems for small molecules, macromolecules and lipids, and it supports various physicochemical states, such as gas, sol, gel, and solid samples, enabling interaction, adsorption, and diffusion analyses. With the development of information technology in recent years, measurement informatics approaches for effectively utilizing accumulated NMR data are becoming more and more important. Because NMR approaches can produce a number of data with high reproducibility and inter-institution compatibility, further analysis of such data using multivariate analysis and machine learning approaches is often worthwhile[5]. However, NMR spectrum analysis of the mixture is difficult due to noise and signal duplication, and requires a lot of labor. Therefore, there is a need for research on data cleansing such as data quality control, noise reduction and signal separation, which is the previous stage of spectrum analysis, and data-driven analysis that makes use of the data accumulated over many years. To overcome challenges in pre-processing and data analysis in NMR, measurement informatics is of academic and social importance (Figure 2).

Part II described a low-resolution NMR signal assignment approach based on the database and pre-processing. The signal assignment is one of the important problems in NMR spectrum analysis. Standard spectra and chemical shift databases are typically used to assign signals. Many chemical shift databases are available on the internet, and using these services is simple and convenient[6]. However, if the database includes candidate molecules that are not present in the sample but have similar partial structures to sample components, false positives may arise. The applicability of this NMR approach has been limited to specific kinds of small molecules and high-resolution NMR measurements due to the molecular complexity and unmatured informatics techniques. In

this study, I examined a signal assignment approach that used pre-processing methods to enhance and separate peaks for a broad NMR spectrum.

Part III described data cleansing approach. In this research, I propose the concept of a data cleansing strategy for improving the problem of data governance and noise. The application of NMR has spread not only to data-driven science at academia but also to quality control in the industry. Effective utilization of NMR data accumulated for many years is important for the measurement informatics field in the materials society. In addition, the value of raw NMR data for transparency, reproducibility, integrity, and reuse of research data increases in recent years[7,8]. The quality of raw data is important because it influences the value of knowledge gained in insight and predicrion. Recently, a database that associates Free Induction Decay (FID) and chemical structure has been proposed[9], but there is no system to associate research papers with original data, promote reuse in data-driven science, and support scientific discovery. Furthermore, the presence of noise in the raw FID is also a problem as the data quality. Previous research focused mainly on increasing the signal intensity or decreasing the noise and many signal processing methods have been proposed to improve the signal-to-noise ratio (SNR)[10]. However, the characteristics of the FID noise and the relationship between the parameters and the noise are not clear. Therefore, I performed noise factor analysis to investigate the relationship between the measurement parameters of NMR data accumulated in the laboratory and other institutions and the signal-to-noise ratio (SNR). I have developed a signal separation method that combines short-time Fourier transform (STFT) and probabilistic sparse matrix factorization (PSMF) for solution NMR signals containing different motility (T_2^* relaxation, which is referred to decay of transverse magnetization caused by a combination of transverse relaxation and magnetic field inhomogeneity) components and noise.

Part IV described the material development approach using solid-state NMR. Solid-state NMR (ssNMR) spectroscopy, especially anisotropic interactions, carries high-order structure and dynamic information of the sample. The high-order structure of materials exerts a significant influence on their macroscopic properties[11]. However, such an analysis is difficult because of the broad and overlapping spectra of these materials. Therefore, the separation of ssNMR signals is an important issue for extracting the information hidden in the NMR spectrum of materials having diversity in high-order structures and components. The application of the signal deconvolution method was examined for ssNMR signals of multi-component materials having multiple domain structures or components in the solid-state. After pre-processing of data, a data mining step such as multivariate analysis is usually performed. The multivariate analysis enables

combined analysis of multiple datasets derived from various measurement conditions and instruments, obtaining new information by interpreting them collectively. Traditional material development approaches are experimentally driven and trial-and-error are facing significant challenges due to the vast design space of materials. To address these problems, machine-learning-assisted materials development is emerging as a promising tool for successful breakthroughs in many areas of science[12]. In this study, the application of Generative Topographic Mapping Regression (GTMR) was examined in order to predict NMR signals and physical properties as descriptors of components and structures.

Based on the above method, I have developed a measurement informatics approach useful for NMR analysis of molecular complex systems.

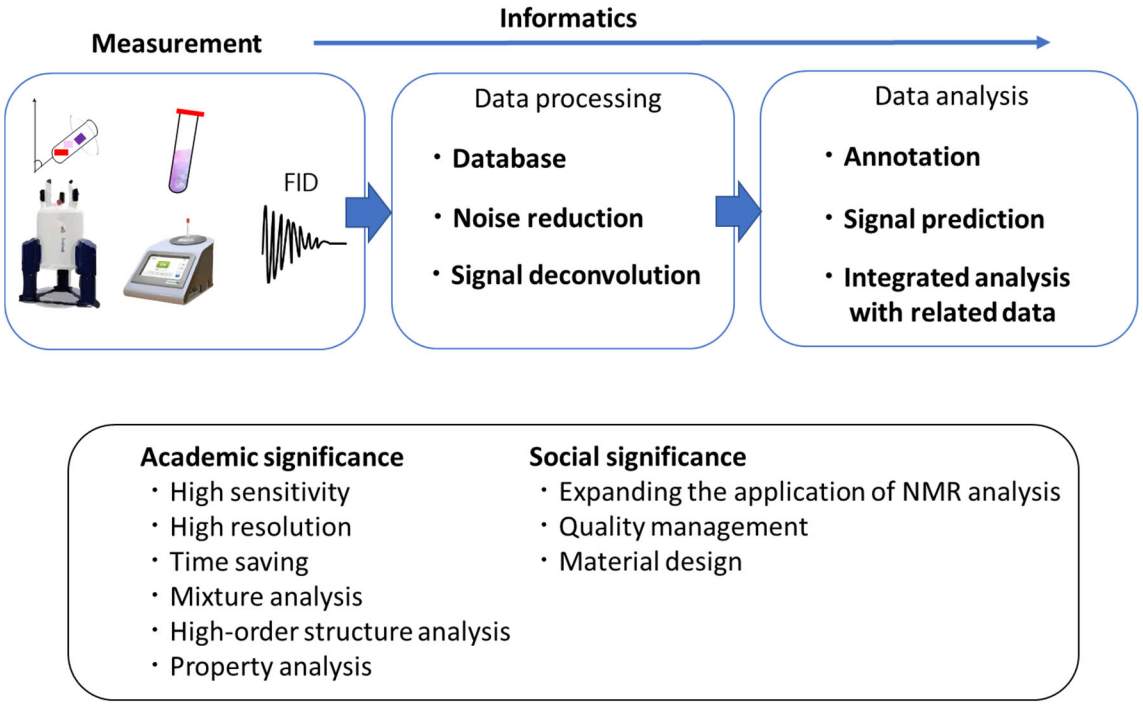


Figure 2. Challenges and significance of measurement informatics in this doctoral research. This figure shows the challenges and significances of data processing and data analysis after NMR measurements in measurement informatics.

Part II

Low-resolution NMR signal assignment approach based on database and pre-processing

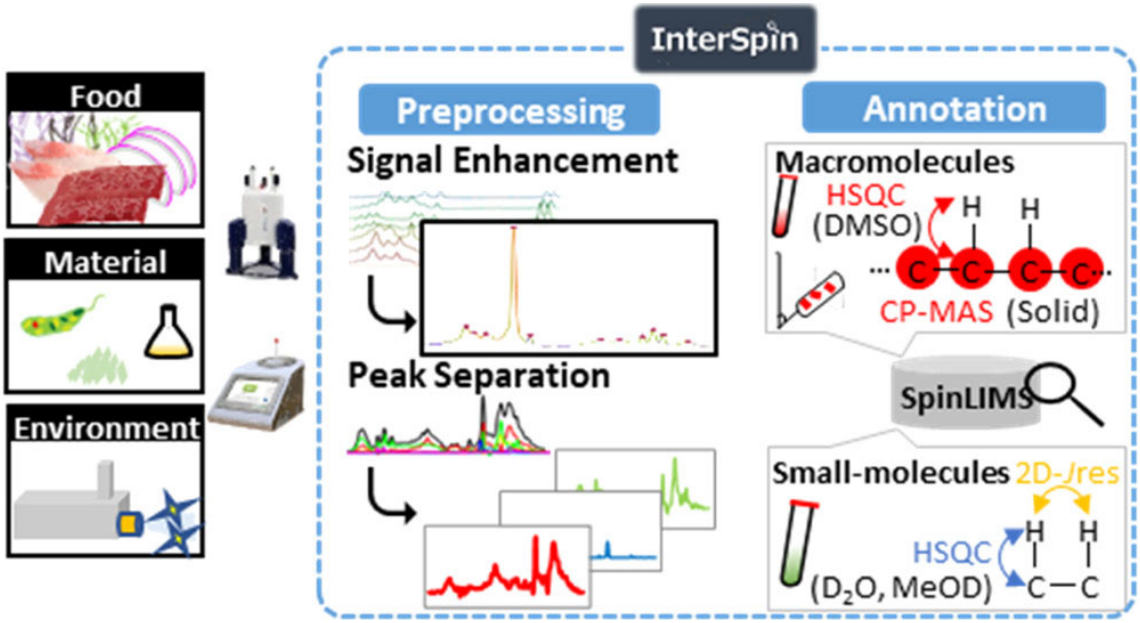
Chapter 2

InterSpin: Integrated Supportive Webtools for Low- and High-Field NMR Analyses Toward Molecular Complexity

2.1 Abstract

InterSpin (http://dmar.riken.jp/interspin/) comprises integrated, supportive, and freely accessible preprocessing webtools, and a database to advance signal assignment in low- and high-field nuclear magnetic resonance (NMR) analyses of molecular complexities ranging from small molecules to macromolecules and lipids for food, material, and environmental applications. To support handling of the broad spectra obtained from solid-state NMR or low-field benchtop NMR, we have developed and evaluated two preprocessing tools: SENSI, which enhances the signal-to-noise ratio by spectral integration; and PKSP, which separates overlapping peaks by several algorithms, such as non-negative sparse coding. In addition, the SpinLIMS database stores numerous standard spectra ranging from small molecules to macromolecules and lipids in solid and solution state (dissolved in polar/nonpolar solvents), and can be searched under various conditions by using the following molecular assignment tools. SpinMacro supports easy assignment of macromolecules and lipids in natural mixtures via solid-state ^{13}C peaks and DMSO-d6 dissolved ^{1}H-^{13}C correlation peaks. InterAnalysis improves the accuracy of molecular assignment by integrated analysis of ^{1}H-^{13}C correlation peaks and ^{1}H-*J* correlation peaks of small molecules dissolved in D_2O or MeOD-d4, supports easy

narrowing down of metabolite candidates. Lastly, by enabling database interoperability, SpinLIMS's client software will ultimately support scientific discovery by facilitating sharing and reusing of NMR data.



Graphical abstract

2.2 Introduction

Environmental problems such as marine pollution, destruction of land and fresh water ecosystems, depletion of resources including energy, raw materials, and food, and health problems are some of the global challenges of modern society. The realization of a materials-circulating society, including use of renewable energy and production of sustainable food and materials, is increasingly important. With the rapid development of information and communication technology in recent years, it is expected that innovations in environmental science, sustainable resources, materials, foods, and medicine, will be integrated by effectively connecting the accumulating scientific data and real-world information[13-15]. As a result, digital innovations in the analyses of natural mixtures, such as biogeochemical samples from the environment and molecular complexities from biological tissues, are becoming important both for a sustainable society and for healthcare[5,6].

Nuclear magnetic resonance (NMR) approaches to natural mixture analysis are being developed as a strategy[16] to evaluate homeostatic stages via molecular

compositional changes in healthcare[17-21], foods[22,23], natural materials[24-27], biomass utilizations[28,29], and environmental ecology[30-34]. Alongside, there have been many advances in NMR technology, including high-field NMR over 1 GHz using high-temperature superconducting materials[35], hyperpolarization[36] and photodetection NMR using diamond nitrogen-vacancy centers[37], zero-magnetic field NMR[38], and compact and benchtop NMR instruments that have become highly cost-effective owing to the marked progress in permanent magnet materials[39,40]. These innovations in NMR hardware are likely to be applied not only to precise analysis by high magnetic field NMR in the laboratory but also to homeostatic assessments of environment and health, and quality control in the fields of agriculture, forestry, and fishery.

Thus, identification of molecules contained in mixtures is an important task in NMR analysis. Because the physical and chemical properties of these molecules can be extremely diverse, various sample preparation methods and pulse sequences have been used for mixture analysis[5,6,41]. Depending on the target molecules under analysis, the sample preparation method may range from solid-state to polar and nonpolar solvent systems[42,43]. When targeting small molecules, for example, solution NMR in a polar or semipolar solvent system such as deuterated water (D_2O) or deuterated methanol (MeOD-d4) is generally used[44,45]. On the other hand, macromolecules can be evaluated by using a dimethyl sulfoxide (DMSO-d6) solubilized system[46] or by solid-state ^{13}C- cross-polarization magic-angle spinning (CP-MAS) NMR. 1D-NMR and 2D-NMR such as ^{1}H-^{13}C hetero-nuclear single quantum coherence (HSQC) and 2D-^{1}H-*J* resolved (2D-*J*res) spectroscopy are also useful for applications where stable isotope labeling experiments cannot be applied.

Nevertheless, such molecular assignments remain difficult owing to the problems of spectral overlap, and a lack of available reference spectra or convenient molecular assignment tools specific to the molecules and conditions of interest. Databases and analytical tools for traditional major metabolomics studies such as HMDB[47], BMRB[48], BML-NMR[49], MMCD[50], NMRShiftDB[51], TOCCATA[52], COLMAR[53], MetaboLights[54], MetaboAnalyst[55], SpinAssign[56], and SpinCouple[57] focus on the analysis of low molecular mass metabolites by high magnetic field solution NMR. For the analysis of macromolecular mixtures derived from environmental samples and living organisms, however, solid-state CP-MAS spectral can characterize insoluble samples, whereas HSQC spectral data in DMSO-d6 solvent are required to characterize soluble samples. The BMRB contains reference NMR data on biomolecules in various solvents such as DMSO-d6 and methanol, but it is limited to partial structural data for polysaccharides. In addition, Bm-Char of ECOMICS[58] can be used to characterize

chemical structures from the HSQC spectrum of a biomass sample. As opposed to many other databases of metabolites, GISSMO[59,60] offers the complete spin system for a large number of metabolites, making analysis possible regardless of the magnetic field. Nevertheless, there remain insufficient databases and analytical tools for complex mixtures of similarly structured macromolecules, or for solid CP-MAS NMR, which has typically very low resolution, or low-field benchtop NMR.

In order to overcome these problems, here we have developed InterSpin, an integrated supportive webtool comprising freely accessible preprocessing tools, a database, and molecular assignment webtools to advance signal assignment in low- and high-field NMR analyses of mixtures from small molecules to macromolecules and lipids (Figure 1). InterSpin comprises the following three elements: 1) spectrum preprocessing tools; 2) molecular assignment tools; and 3) the SpinLIMS (InterSpin Laboratory Information Management System) database.

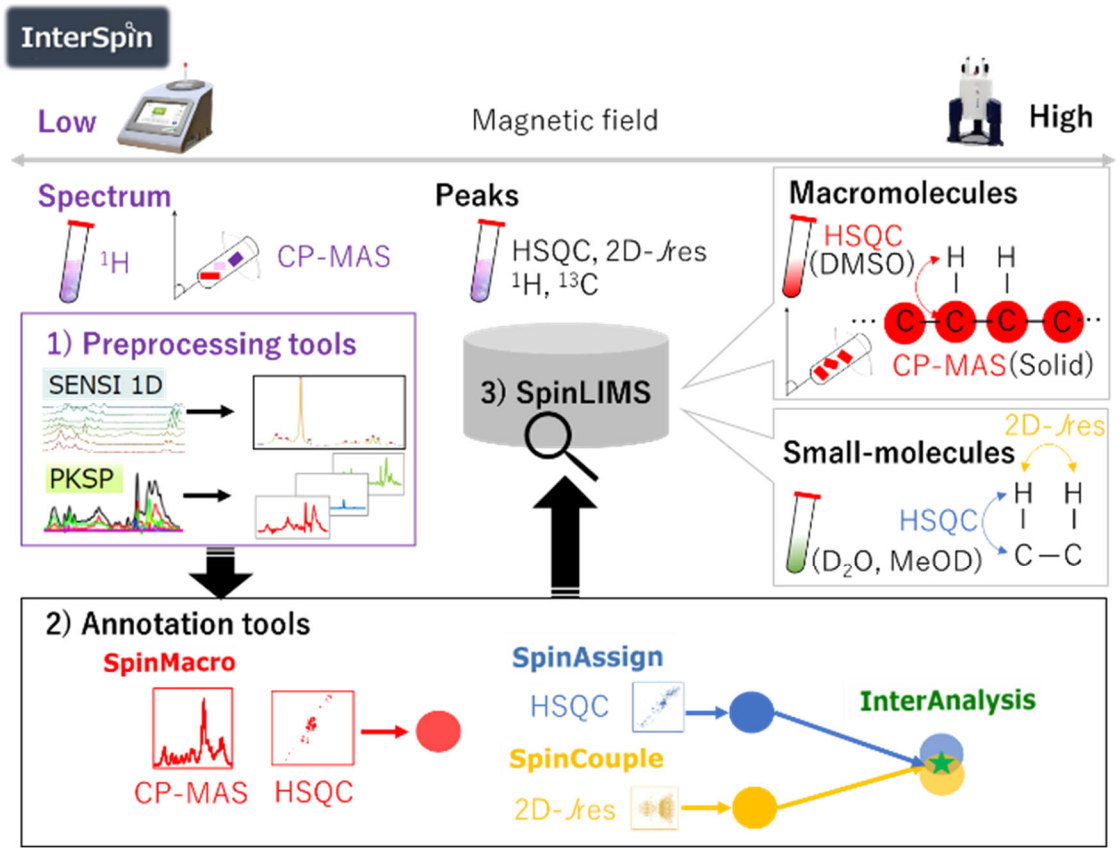


Figure 1. Overview of InterSpin. InterSpin is a freely accessible integrated supportive webtool for advanced performance of NMR signal assignment in low- and high-field NMR analysis of mixtures from small molecules to macromolecules and lipids. InterSpin comprises the following three elements. 1) Spectrum preprocessing tools. In the case of a broad spectrum obtained from low-field benchtop ^{1}H-NMR or solid-state ^{13}C-CP-MAS, SENSI (SENsitivity improvement with Spectral Integration) helps to overcome the problem of low signal-to-noise ratio by increasing resolution through the

integration of multiple spectra, while PKSP supports effective peak separation by a multivariate spectral decomposition method. 2) Molecular assignment tools. SpinMacro supports simplifying the assignment of a solid CP-MAS spectrum or a DMSO-d6 solubilized ^{1}H-^{13}C HSQC spectrum for macromolecules and lipids. SpinAssign searches the SpinLIMS database for a compound corresponding to the HSQC NMR peaks. SpinCouple can assign ^{1}H-*J* 2D-*J*res NMR peaks. InterAnalysis is a Venn diagram-type highly accurate annotation tool that helps to narrow down candidate molecules by using correlation peaks from both the HSQC spectrum and the 2D-*J*res spectrum. In the bottom right of the figure, blue, yellow, and red circles represent a set of search results; the green star represents the narrowed down set. 3) SpinLIMS (InterSpin Laboratory Information Management System) database. The database includes reference solid-state CP-MAS spectra and solution-state HSQC spectra (DMSO-d6) for macromolecules and lipids, and reference solution-state HSQC and 2D-*J*res spectra (D_2O and MeOD-d4) for small molecules.

2.3 Results and Discussion

3.2.1 Signal enhancement and peak separation of benchtop NMR spectra by SENSI and PKSP

To support preprocessing of a broad spectrum, InterSpin uses PKSP (PeaKs SeParation) and SENSI1D, which have been newly developed as webtools (Figure 1). SENSI1D is intended to increase signal intensities and to overcome the problem of low signal-to-noise (S/N) ratio by the integration of multiple spectra without additional measurements. On the other hand, PKSP is a multivariate method of spectral decomposition that includes the algorithms non-negative sparse coding (NNSC)[61,62], which separates the spectrum into non-negative sparse components; multivariate curve resolution – alternate least squares (MCR-ALS); fast independent component analysis (Fast ICA); and non-negative matrix factorization (NMF). We have previously described the spectrum-preprocessing methods of SENSI[63], MCR-ALS[25], and NMF[26]; here, we have integrated them into InterSpin as a freely available webtool.

First, we verified the effectiveness of the new function NNSC in PKSP by using multicomponent test data with increasing numbers of components. MCR-ALS and NMF required significant computing time when processing more than 100 components, whereas NNSC and Fast ICA were fast, maintaining speed even as the component number increased (Figure 2). In terms of resolving the spectrum of mixtures of 10 standard compounds (Supporting Information Table S1) with reference to the spectrum of each standard compound by PKSP, the Durbin-Watson (DW) plot approached 2 (Supporting Information Figure S1, white) with 10 components identified by all algorithms, the residual sum of squares (RSS) plot converged to 0, and the spectrum was separated into the correct number of components (Supporting Information Figure S1). NNSC, MCR-ALS and NMF generally showed good separation of all components from the mixed

spectra (Supporting Information Figure S2). In NNSC, a sharp peak was observed in the broad part of the spectrum (3 to 4 ppm) for glucose. In Fast ICA, a large error in the original spectrum occurred for glucose and sucrose. In NNSC, NMF, and MCR-ALS, the ratio of components in the mixture was well estimated, but Fast ICA showed an error for alanine, phenylalanine, proline, valine, and glucose, although its calculation speed was fast (Supporting Information Figure S3).

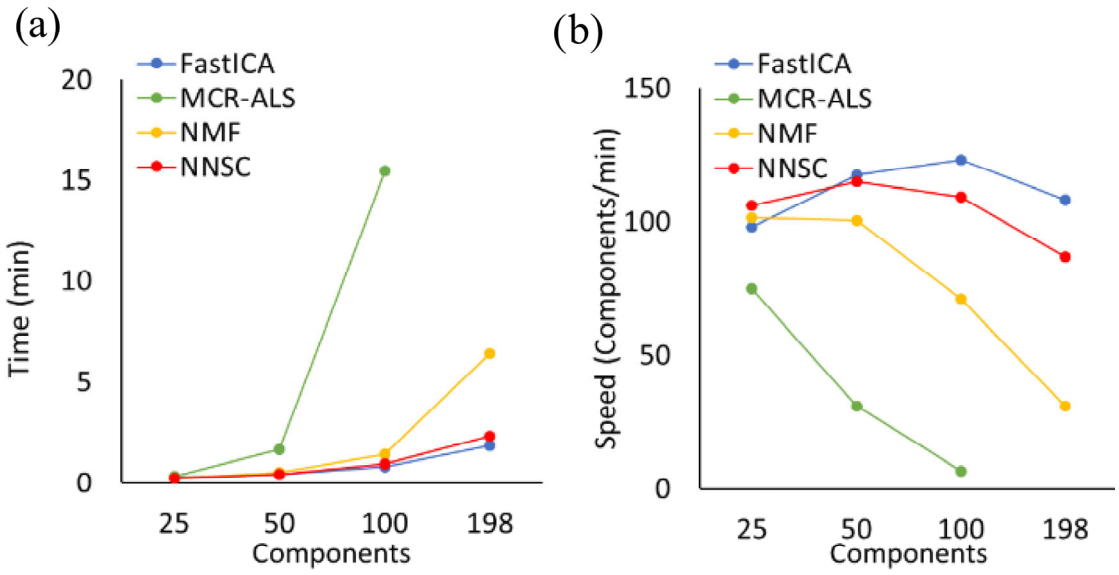


Figure 2. Comparison of the analysis speed of each algorithm in PKSP (PeaKs SeParation). (a) Three average analysis times for 25, 50, 100, and 198 components (i.e., compounds to be separated by each algorithm of PKSP). (b) Three average analysis speeds for 25, 50, 100, and 198 components.

As a demonstration of the integrated use of SENSI and PKSP webtools, Figure 3 shows that histidine, creatine, and lactate were well separated as major components of Thunnus muscle measured by benchtop 60-MHz NMR (Figure 3). For this demonstration, the 60-MHz NMR spectra from 51 samples of 40 fish foods (Supporting Information Table S2) and 11 standard compounds (Supporting Information Table S3) first showed that the SENSI tool strengthened 25 peaks of the 11 standards 66-fold on average and improved the S/N ratio 5.5-fold (Supporting Information Figure S4, Table S4). Subsequently, peak separation of the benchtop NMR spectrum was performed by NNSC of PKSP, which led to the separation of 17 components (Supporting Information Figure S5). Note that where there are multiple signals for the same molecule, their coefficients of variation (CVs) indicate that their signal intensities vary together. This information can support signal attribution. Thus, histidine, creatine, and lactate could be identified by

using the CV value of peaks detected by SENSI (Supporting Information Figure S4b) and the individual components obtained by PKSP (Supporting Information Figure S5d).

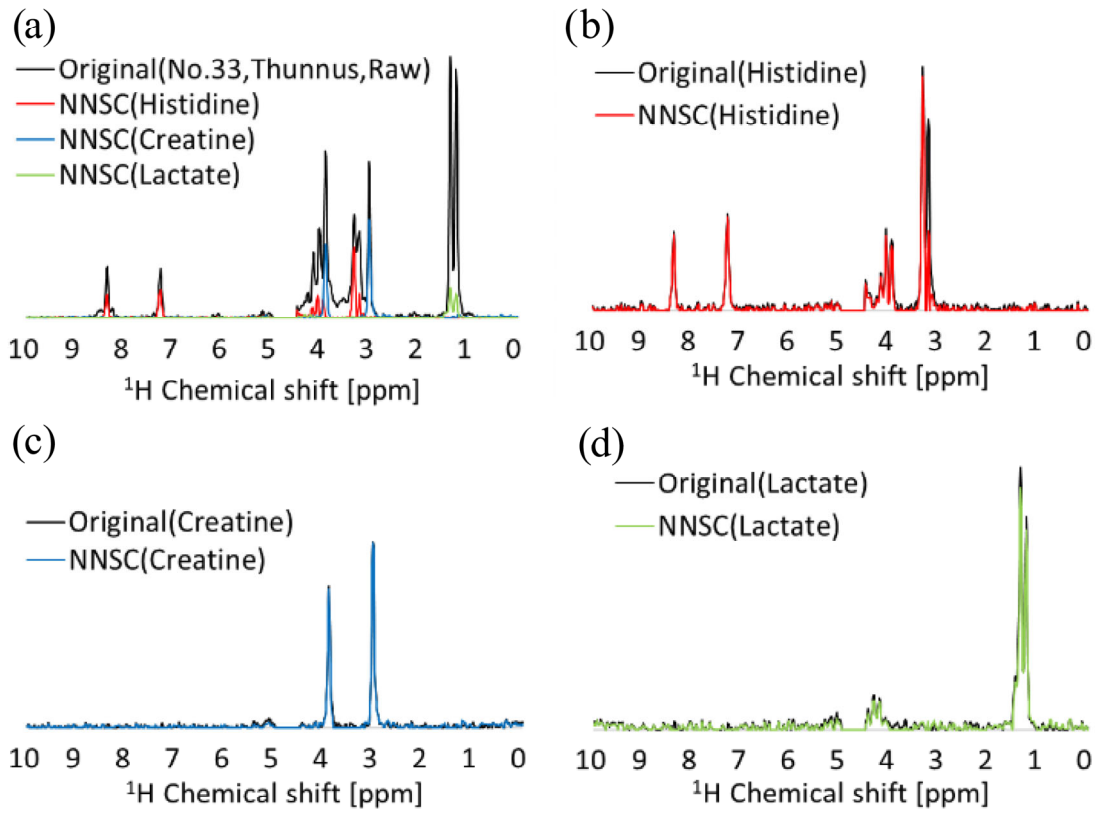


Figure 3. Molecular assignment of a mixture using peaks separated by PKSP (NNSC). (a) Original and separated spectra of No. 33 Thunnus sample measured by benchtop 60-MHz NMR. (b–d) Original and separated component spectra of histidine (b), creatine (c), and lactate (d).

To evaluate peak separation by the four algorithms in PKSP, here we determined the appropriate number of separate peaks using DW and an RSS plot, which is the sum of squares of the residuals of the original matrix of each model and the reconstruction matrix[64]. Although FastICA calculated negative values as separate matrices, it determined the number of components more quickly than the other algorithms (Figure 2). When the number of components was large, however, NNSC provided a realistic approximate spectrum at high speed and with nonnegative values. For the analysis of large numbers of components, therefore, we considered that it would be most efficient to determine the number of components with Fast ICA and then perform accurate spectral separation with NNSC.

For analysis in SENSI and PKSP, the peak maximum must have exactly the same chemical shift for each signal. Thus, peak alignment to correct chemical shifts altered due to pH, temperature, or magnetic field inhomogeneity caused by magnetic material in the sample is an important process.

The calculation algorithm used for PKSP is multivariate analysis; therefore, it is essential to have M numbers of spectral data. However, because 2D-NMR has data in the f2 direction, PKSP can be applied to data from the 2D matrix of one or more spectra of 2D-NMR. Therefore, if a user has difficulty with 2D-NMR peak picking of, for example, saccharides and lipids, our approach can support objective peak picking by helping to separate peaks via PKSP. As a result of peak separation in one spectrum of 2D-*J*res using PKSP's MCR-ALS algorithm, it was separated into three components (Supporting Information Figure S6).

In general, quality control is essential in modern food production. In many cases, however, the primary production or distribution sites (i.e., farms or fishing grounds) are located far from laboratories or analytical centers (i.e., food companies or facilities). In such cases, benchtop NMR may potentially revolutionize the quality control processes that identify metabolic changes in food resulting from storage and fermentation. As a practical tool, we previously developed FoodPro[65], a database and webtool for predicting the taste and longevity of foods based on the similarity of desktop NMR spectra of food substances. As shown in this study, SENSI and PKSP are expected to lead to improved cost-effectiveness of this approach by supporting the annotation of the broad spectra obtained from *in situ* low-field NMR.

3.2.2 Assignment of macromolecules and lipids by SpinMacro

InterSpin's Annotation tools consist of the newly developed SpinMacro and InterAnalisys, and the re-implemented SpinAssign and SpinCouple, which were previously developed (Figure 1). SpinMacro is a webtool for supporting simplification of the molecular assignment of macromolecules and lipids in solid-state ^{13}C CP-MAS spectra and in ^{1}H-^{13}C HSQC spectra recorded in DMSO-d6 solvent (Figure 4, Supporting Information Figure S7). As reference data for SpinMacro, solid CP-MAS peaks and HSQC peaks of compounds in DMSO-d6 solvent have been stored in the SpinLIMS (InterSpin Laboratory Information Management System; see Figure 1) database.

The steps for using SpinMacro are as follows. 1) PHP interpretation of the user query for CP-MAS peaks or HSQC peaks. 2) Connect to the SpinLIMS database and

search for candidate molecules within the set range of ^{13}C chemical shifts for CP-MAS, or ^{1}H and ^{13}C chemical shifts for HSQC. 3) Conversion of results to HTML and JavaScript for convenient and quick display. Here, the previously reported solid-state CP-MAS spectra of *Euglena gracilis*[24] and standards (paramylon, peptides, lipids) were queried using SpinMacro and SENSI-PKSP. First, the CV value was determined for peaks picked by SENSI (Figure 5) and then the components were identified by PKSP (Supporting Information Figure S8). As a result, paramylon, peptides, lipids were separated as the main three components of *E. gracilis*. Ultimately, as a result of retrieving the peaks picked by SENSI with SpinMacro, it was possible to verify their assignment (Supporting Information Figure S7). In a previous study of general lipids and general peptides of *E. gracilis*[66], we conducted experiments that required considerable measurement time, such as 2D-/ 3D-NMR pulse sequences of solid-state NMR (i.e., INADEQUATE, SHA+ and 3D-DARR). Because the peak separation by NNSC corresponds to 1D-CP-MAS, this tool supports a more rapid evaluation of mixtures including macromolecules and lipids.

The database and mixture analysis tool for macromolecules and lipids and solid CP-MAS NMR of complex and similar structures have room for development. SpinMacro developed herein retrieves the peak of the whole structures about macromolecules and lipids from SpinLIMS and provides candidate molecules in analyses of environmental and biological macromolecules and lipids. In the future, it should be possible to improve assignment accuracy by discriminating macromolecules and lipids with similar structures through the extraction of features of chemical structures using machine learning algorithms based on databases of macromolecules and lipids.

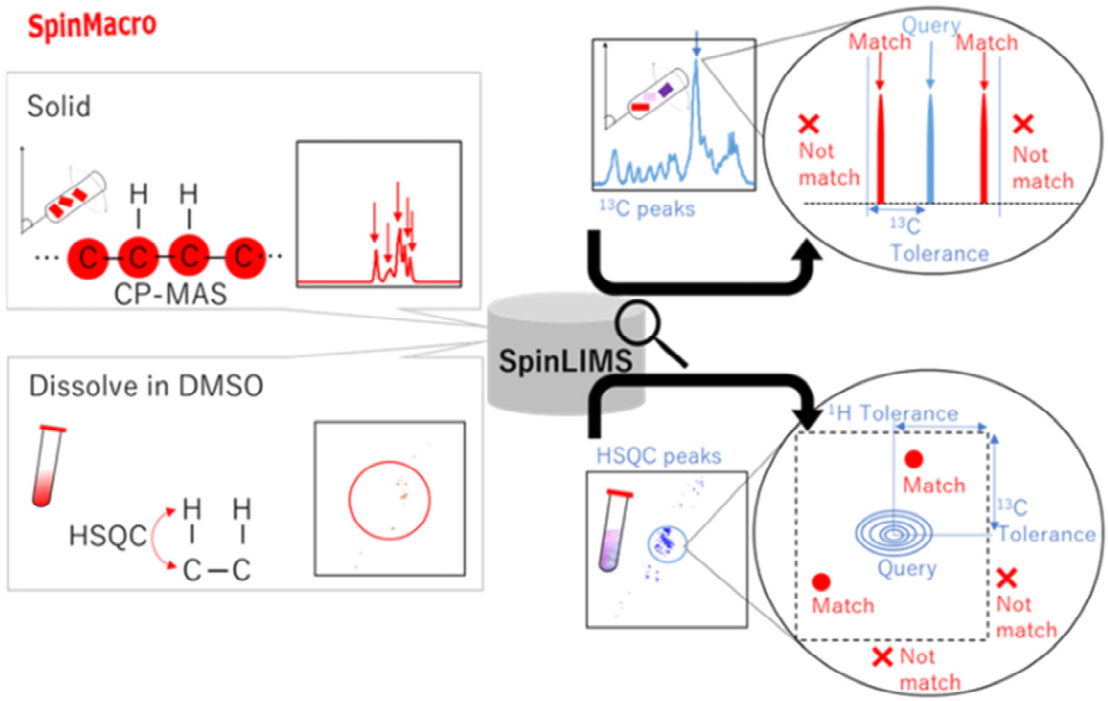


Figure 4. How to assign macromolecules and lipids in a mixture using "SpinMacro". Shown is the flow of data through SpinMacro. The user query of CP-MAS peaks or HSQC peaks are entered as PHP. The SpinLIMS database is then searched for candidate molecules within the set range of ^{13}C chemical shifts for CP-MAS, or ^{1}H and ^{13}C chemical shifts for HSQC.

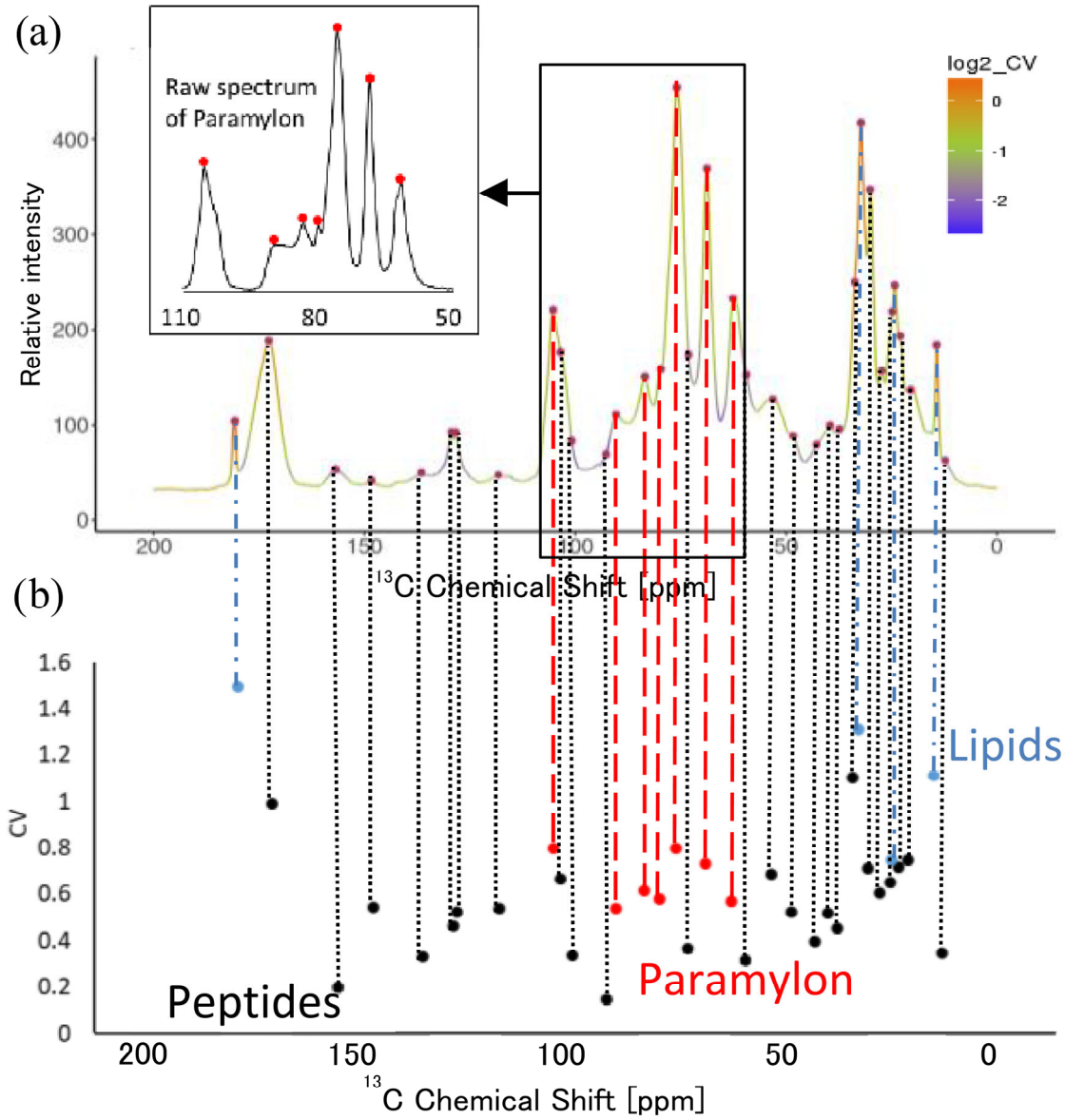


Figure 5. CV of peaks picked by SENSI from *E. gracilis* CP-MAS spectrum. (a) SENSI results. Red circles are picked peaks. The enlarged view (top left) shows the raw spectrum of paramylon from the data used for SENSI of sugar region. (b) CV of peaks picked by SENSI. Blue circles indicate lipids signals, black circles indicate peptides signals, and red circles indicate paramylon signals.

3.2.3 SpinLIMS (InterSpin Laboratory Information Management System) database

Within InterSpin, SpinLIMS (Figure 1) is a relational database comprising several entities or "tables" developed by MySQL (Supporting Information Figs. S9a and S10a, Core tables). To make the database extensible, SpinLIMS client software was developed to incorporate a simple registration system. After registering in the user table ("limsuser"), the researcher can associate their NMR spectrum ("spectrum") with the chemical shift ("cs") or the *J* value ("jval") tables, as well as the molecular value ("metabolite") table by means of the assignment table (cs_assign, hc_pk (h_pk for ^{1}H-1D NMR, c_pk for ^{13}C-1D NMR), hj_pk) via the client software. For the NMR spectrum, there is an associated pulse type table ("pulse"), solvent table ("solvent"), standard substance table ("stdref"). For chemical shifts and *J* values, there is an associated peak shape table ("pkshape"). Molecular name ("metabolitename"), atom ("atom"), nuclide ("nucleus") tables are associated with the molecule.

SpinLIMS contains numerous reference spectra of small molecules to macromolecules and lipids recorded in solid state and solution state (polar and nonpolar solvent systems) that can be used to support mixture analysis of various samples. Overall, there are 34 data tables in SpinLIMS, as well as tables for managing the information from NMR experiments (Supporting Information Figure S10b). In addition to HSQC in D_2O (705 spectra) and 2D-*J*res in D_2O (623 spectra), SpinLIMS has several newly added spectra from CP-MAS (35 spectra), HSQC in MeOD-d4 (947 spectra) and deuterated DMSO-d6 (171 spectra), and 2D-*J*res in MeOD-d4 (357 spectra). SpinMacro, InterAnalysis, SpinAssign, and SpinCouple are connected to the MySQL server via a local network in InterSpin (Supporting Information Figure S9b). As a result, the re-implemented SpinAssign and SpinCouple facilitate chemical shift searches in MeOD-d4. SpinAssign also facilitates searches in deuterated DMSO-d6/pyridine-d5 solvent.

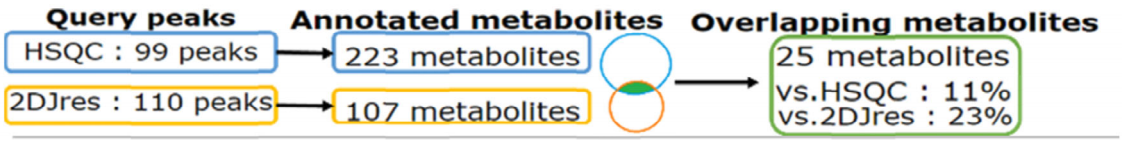
3.2.4 Venn diagram-type annotation by InterAnalysis

Within InterSpin, the new tool InterAnalysis is a Venn diagram-type annotation tool that can aid simultaneously searches of two kinds of correlation peak, ^{1}H-^{13}C HSQC and 2D-*J*res, to narrow down candidate molecules (Figure 6). The flow of data through InterAnalysis is as follows: 1) PHP interpretation of user queries; 2) connection to the

SpinLIMS database and conversion to HTML; and 3) JavaScript execution for a convenient and rapid view.

Here, we demonstrated the application of InterAnalysis to HSQC and 2D *J*res peaks from *Acanthogobius flavimanus* (Yellowfin goby) body muscle extracts in MeOD-d4 (Figure 6) and deuterated potassium phosphate (Supporting Information Figure S11). For data acquired in MeOD-d4 extract, SpinAssign and SpinCouple assigned 223 and 107 molecules. By contrast, InterAnalysis assigned 25 molecules, narrowing down the molecules to 11% and 23%, respectively (Figure 6). From previous studies[25,31,33], seven metabolites such as L-valine, L-leucine, L-phenylalanine, L-histidine, L-proline, Linoleic acid and Capric acid were confirmed as well-known metabolites that should be present in fish.

In the analysis of natural mixtures, molecular assignment based on two kinds of 2D-NMR spectra, HSQC and 2D-*J*res, is a powerful strategy to increase assignment accuracy. The previous tools, SpinAssign and SpinCouple, acquired two separate results of correlation peak attribution; thus, it was highly time-consuming to narrow down candidate molecules. The newly developed Venn diagram-type webtool, InterAnalysis, supports the annotation of environmentally and biologically derived small-molecule mixtures.



Input:HSQC 1h ppm	Input:HSQC 13c ppm	Input:2DJ 1h ppm	Input:2DJ J val hz	Annotated metabolite name
1.0119	19.1922	2.2448	-1.5403	cholic acid
1.0119	19.1922	2.2448	-1.5403	L-valine
1.0119	19.1922	1.0316	-3.5008	L-valine
1.0119	19.1922	0.8779	7.4216	L-valine
1.0119	19.1922	1.0549	3.5008	L-valine
1.0119	19.1922	2.2341	0	cholic acid
1.0119	19.1922	3.3608	0	cholic acid
0.9845	24.7686	3.5184	-2.7306	L-leucine
0.9845	24.7686	0.9752	2.7306	L-leucine
0.9845	24.7686	3.5106	2.3805	L-leucine

Showing 1 to 10 of 486 entries

Figure 6. Result of InterAnalysis for ^{1}H-^{13}C HSQC and ^{1}H-*J* 2D *J*res peaks from *Acanthogobius flavimanus* body muscle extract in MeOD-d4. Summary shows the number of query peaks, the number of assigned molecules, and the narrowed down set of molecules. The table shows some of the molecular assignment results for each query peak.

2.4 Materials and Methods

2.4.1 SENSI and PKSP

The SENSI and PKSP webtools were developed using the Shiny package based on previously reported R scripts[25,26,63]. Here, we incorporated a new method, NNSC[61,62], into PKSP.

2.4.2 Database and client software of SpinLIMS

SpinLIMS was developed in MySQL. It integrated previously reported data from SpinAssign[56] and SpinCouple[57]. In addition, it newly implemented NMR spectra for solid-state CP-MAS and solution-state in DMSO-d6/pyridine-d5 and MeOD-d4 solvents. The SpinLIMS client software was developed with Java.

2.4.3 SpinMacro, InterAnalysis, SpinAssign, and SpinCouple

SpinMacro and InterAnalysis were developed in HTML, PHP, JavaScript, and MySQL. SpinAssign[56] and SpinCouple[57] were completely re-implemented within the program and were connected to the SpinLIMS database to run within InterSpin.

2.4.4 Evaluation of benchtop NMR signal assignment performance by SENSI and PKSP

To evaluate the performance of SENSI and PKSP, ^{1}H-NMR data of 40 fish-based food mixtures and 22 standard compounds measured by benchtop NMR at 60 MHz were conducted phase and baseline correction, spectral alignment and normalization and then analyzed. In addition, the analysis speed of PKSP was evaluated by using similarly processed the 500 MHz ^{13}C-CP-MAS spectrum of plant and algae biomass with 198 components.

2.4.5 Evaluation of assignment about macromolecules and lipids by SpinMacro with SENSI and PKSP

To evaluate the molecular attribution strategy of SpinMacro using SENSI and PKSP, we used the previously reported CP-MAS spectrum of *E. gracilis*[24] and spectra of standard compounds (paramylon, peptides, and lipids). NMR spectrum were conducted phase and baseline correction, spectral alignment and normalization and then analyzed.

2.4.6 Comparison of small-molecule assignment by InterAnalysis, SpinAssign, SpinCouple

To evaluate the performance of InterAnalysis, HSQC and 2D-*J*res peaks in a 700 MHz NMR spectrum of body muscle extract of *Acanthogobius flavimanus* were assigned molecules by InterAnalysis, SpinAssign, and SpinCouple.

2.5 Conclusions

As shown above, InterSpin provides free access to a suite of tools whose goal is to support the interpretation of low-resolution NMR spectra, similar to the spectra recorded for food, material, and environmental applications. Each tool of InterSpin supports low-resolution NMR spectrum analysis by having interoperabilty as demonstrated, for example, by the peak attribution of *E. gracilis* by NNSC of SENSI, and confirmation of metabolite candidates by SpinMacro. Fur-thermore, 2D-*J*res and HSQC are pulse sequences that are frequently used in high-magnetic field NMR; conventionally, SpinAssign and SpinCouple have had to be applied individually, but InterAnalysis will aid the simultaneous application of these tools at the same time.

NMR has the great advantage that chemical shifts and coupling constants are absolute physical constants that have high repeatability and interchangeability between different agencies. Therefore, NMR provides data that are suitable for reuse globally. For NMR analyses that target the molecular complexity of living bodies and environments, InterSpin provides an integrated supportive resource, consisting of an extensible SpinLIMS database and webtools that are easily accessible to varied and numerous researchers. SpinLIMS's client software will ultimately promote scientific discovery through the open circulation of knowledge by facilitating data sharing and reusing, as well as the interoperability of NMR data for the achievements of re-searchers to be recognized fairly and with transparency.

In conclusion, InterSpin comprises integrated supportive webtools that are effective not only for precision analysis in laboratories but also for on-site analysis by benchtop NMR. As a platform linking the laboratory and the real world, it will support sustainable development based on NMR data.

Part III

Data cleansing approach

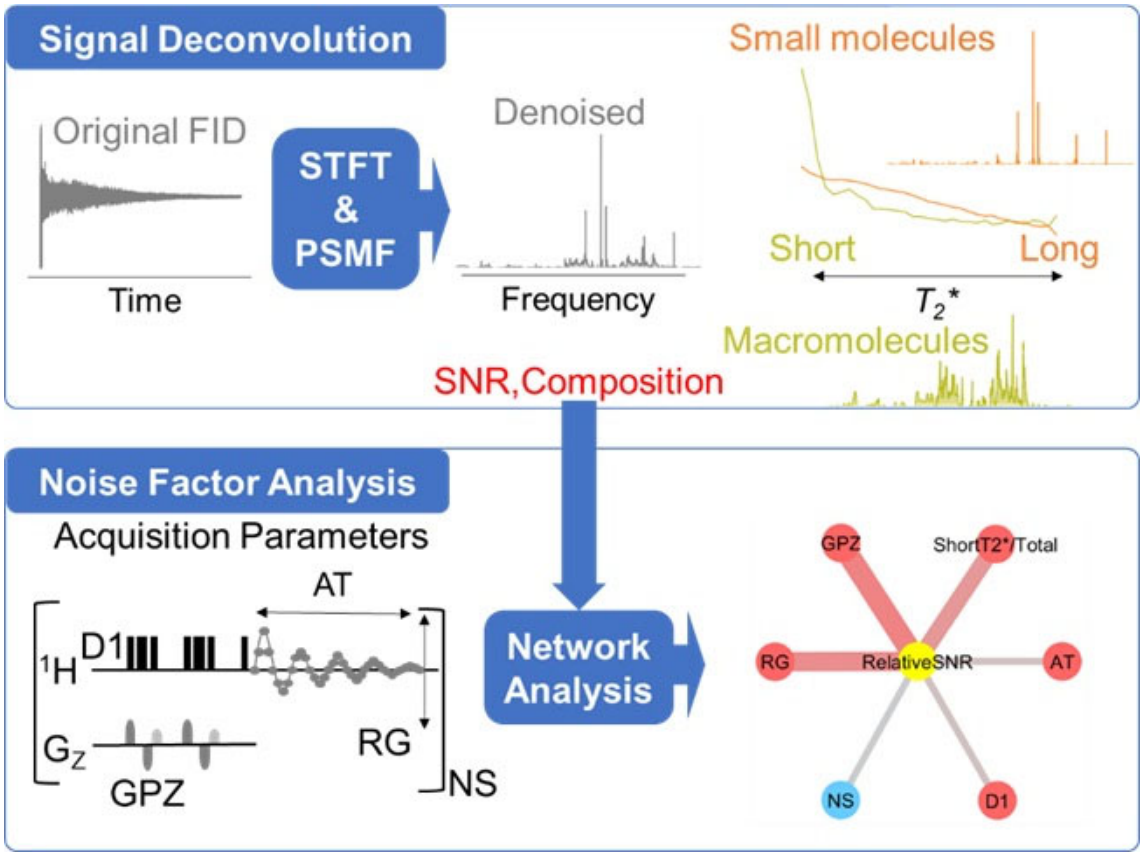
Chapter 3

Signal Deconvolution and Noise Factor Analysis Based on a Combination of Time–Frequency Analysis and Probabilistic Sparse Matrix Factorization

3.1 Abstruct

Nuclear magnetic resonance (NMR) spectroscopy is commonly used to characterize molecular complexity because it produces informative atomic-resolution data on the chemical structure and molecular mobility of samples non-invasively by means of various acquisition parameters and pulse programs. However, analyzing the accumulated NMR data of mixtures is challenging due to noise and signal overlap. Therefore, data-cleansing steps, such as quality checking, noise reduction, and signal deconvolution, are important processes before spectrum analysis. Here, we have developed an NMR measurement informatics tool for data cleansing (freely available at http://dmar.riken.jp/NMRinformatics/) that combines short-time Fourier transform (STFT; a time–frequency analytical method) and probabilistic sparse matrix factorization (PSMF) for signal deconvolution and noise factor analysis. Our tool can be applied to the original free induction decay (FID) signals of a one-dimensional NMR spectrum. We show that the signal deconvolution method reduces the noise of FID signals, increasing the signal-to-noise ratio (SNR) about tenfold, and its application to diffusion-edited spectra allows signals of macromolecules, lipids and unsuppressed small molecules to be

separated by the length of the T_2* relaxation time. Noise factor analysis of NMR datasets identified correlations between SNR and acquisition parameters, identifying major experimental factors that can lower SNR.



Graphical abstract

3.2 Introduction

NMR spectroscopy is one of the most powerful tools available for molecular characterization at the atomic level[67]. Because it is non-invasive, NMR has been applied to data-driven analyses of molecular complexity in many areas of health[68], food[65], materials[69], and the environment[5]. In measuring NMR signals, the main challenges are the sensitivity and resolution of the NMR spectrum[70]. On the one hand, various techniques and devices for improving sensitivity have been developed, such as high-field magnets[71], cryogenic detection systems[72], shimming and locking to adjust the magnetic field[73], and dynamic nuclear polarization[74]. In addition, pulsed field gradient (PFG), nonuniform sampling[75] and magnetization transfer techniques such as cross-polarization[76] and INEPT (Insensitive nuclei enhanced by polarization transfer)[77] have been developed to enhance the sensitivity per unit time. On the other hand, compact

and benchtop NMR instruments with lower resolution have become highly cost-effective owing to marked progress in the materials used for the permanent magnet[78].

Regarding spectral resolution, many pulse sequences for the measurement of one-dimensional (1D)-NMR with selective signal suppression, including pre-saturation, Carr–Purcell–Meiboom–Gill (CPMG)[79], WATER suppression by GrAdient Tailored Excitation (WATERGATE)[80], diffusion-editing[81], double quantum filter[82], and pure shift NMR[83], have been developed to reduce signal overlap. However, the spectra have remaining overlapping signals, or the overlapping peaks themselves contain part of the information of the sample. In this regard, overlapping signals can be separated by two-dimensional (2D)-NMR, in which multiple free induction decays (FIDs) are measured over a small change in evolution time, but this approach is time consuming[84].

Conventionally, methods for improving the sensitivity and resolution of FIDs are adjusted by pre-processing steps, such as zero filling and apodization, before Fourier transformation (FT) is carried out[85]. Other methods for reducing mathematical noise from FID signals focus on the region of interest (ROI), such as reference deconvolution[86], harmonic inversion noise removal (HINR)[87], and complete reduction to amplitude frequency table (CRAFT)[88]. In addition, STFT and wavelet transform[89] have been developed as alternative transformation methods to FT for analyzing the relationship between the time and frequency of FIDs. In principle, the exponential decay constant of the FID obtained by applying a 90° pulse to create transverse magnetization is the T_2 relaxation time, a physical parameter independent of field inhomogeneity. In reality, however, because of the effect of magnetic field inhomogeneity, the decay constant of the FID is defined as T_2*, an instrument-dependent parameter, rather than T_2. STFT has the ability to extract time-varying behavior from FIDs, allowing for the analysis of dynamic chemical shifts of atoms in flexible proteins[90]. In addition, it has been reported that STFT can extract T_2* information from FIDs and improve the results of discriminant analysis[91]. Applying the same idea to covariance NMR[92], T_2*-weighted covariance NMR improves the sensitivity and resolution of signals based on the difference in T_2*, determined by dividing each FID in the t_1 dimension of 2D-NMR to create a series of sub-FIDs[93]. In an alternative approach, matrix factorization (MF) is commonly used to extract signal components and separate peaks in spectra[94]. For example, a noise reduction method using principal component analysis (PCA), which is one of the most commonly used multivariate analysis methods for extracting features of data, has been applied to solid CP-MAS NMR data measured by various parameters[95]. Therefore, the quality and amount of information from FIDs can be maximized by applying corrections based on different characteristics. Nevertheless, all these methods require multiple FIDs obtained by adding either spectral dimensions or multiple conditions of samples or parameters. There is also a computational approach such as CORE (COmponent-REsolved; a multi-component spectral separation approach previously introduced method). It focuses on diffusion coefficients to separate the NMR signals of different compounds in PFG-NMR[96-98]. However, this technique requires a specific NMR probe with a coil for generating PFG.

In the current move toward a digital innovation society, tools for NMR measurement informatics are becoming increasingly important[6]. Alongside this, the value of raw NMR datasets for reuse in research studies is rising[9]. Although the quality of raw data influences the value of knowledge obtained in terms of both insight and prediction[99], data cleansing methods for utilizing various kinds of NMR data

accumulated over many years, such as data quality checks, noise reduction, and signal deconvolution, have not been established.

In this study, by focusing on acquisition parameters[100-104] and noise[89], we have developed an NMR measurement informatics tool for data cleansing based on FID signal deconvolution and noise factor analysis. Our method for deconvoluting signals and noise factor analysis can be applied to original single FIDs from 1D-NMR and is based on STFT[105] and PMSF[106]. It differs from conventional noise reduction using multivariate analysis[98] because it does not require multiple 1D-NMR data that are measured on many samples or acquired with several acquisition parameters. The difference in T_2* on the time axis determined by performing STFT for each frequency component is useful to separate signals based on MF instead of ROI[86-88]. Our method that focuses on the relaxation time utilizes the attenuation behavior of the FID signal without any hardware upgrade for NMR research field. Lastly, we have developed a function for collecting acquisition parameters as a measurement of experimental factors from a directory of NMR data, and investigated the relationship between signal-to-noise ratio (SNR) and acquisition parameters. A researcher performing NMR must select parameters for each experiment, and normally chooses a reasonable set of parameters based on their experience. We show that these parameters can be characterized in terms of their correlation with SNR by a statistical analysis of accumulated NMR datasets. Therefore, this method will be useful to determine the optimal conditions of acquisition parameters.

3.3 Results and Discussion

4.2.1 Signal deconvolution method

In this study, signal deconvolution, based on the combined method of STFT and PSMF, was applied to FIDs of 1D-NMR to separate the components and improve SNR. The theory behind the signal deconvolution method is described in detail in the Supplementary Material. In brief, in FT NMR spectroscopy, the FID is the NMR signal generated by non-equilibrium nuclear spin magnetization precessing along the magnetic field. In general, this non-equilibrium magnetization can be generated by applying a pulse of resonant radiofrequency close to the Larmor frequency of the nuclear spins of the sample. Each FID is commonly a sum of multiple decayed oscillatory signals. These signals return to equilibrium at different rates or relaxation time constants. Thus, analysis of the relaxation times of an FID for a sample gives significant insight into the chemical composition, structure, and mobility of the sample. FIDs acquired by NMR measurement are composed of many signals derived from the sample, in addition to several types of noise, such as external noise, physical vibration, power supply, and internal noise from the spectrometer due to thermal noise. Therefore, an FID can be modeled as:

$$S(t) = S_{signal}(t) + S_{noise}(t) \quad (1)$$

where $S(t)$ is the measured signal, and $S_{\mathrm{signal}}(t)$ and $S_{\mathrm{noise}}(t)$ are sets of ideal signals and signals from different types of noise, respectively (Equation (1) and Supplementary Equation (S1))[107]. The relaxation process can then be described as the exponential decay of the transverse magnetization $S(t)$ (Supplementary Equation (S2))[108]. The shorter the relaxation time T_2*, the more rapid the decay. If an FID has more than one

component, it will be the sum of contributions from each component (Supplementary Equation (S3)).

Whereas standard FT (Supplementary Equation (S4)) contains only the frequency domain, STFT contains both frequency and time domains. Because the FID signal decays exponentially with time, for STFT, it needs to be divided into several small time intervals (segments) to analyze the time–frequency feature accurately, and FT is used to determine the frequency feature of each segment, thereby increasing the accuracy of signal feature extraction. STFT uses a window function to obtain each weighted segment on the time axis, and then applies FT to each segment. STFT of *S(t)* can be written as:

$$STFT_S(\tau,\omega) = \int_{-\infty}^{\infty} S(t)g(t-\tau)\exp(-i\omega t)dt \quad (2)$$

where the window function g is first used to intercept the progress of FT on $S(t)$ around $t=\tau$ locally, and then FT of the segment is performed on t (Equation (2) and Supplementary Equation (S5))[105]. By moving the center position of the window function g sequentially, all the FTs at different times can be obtained.

$STFT_S(\tau,\omega)$ is a complex-valued function (Supplementary Equations (S6–9)) composed of two types of signal: real (*Re*, Supplementary Equation (S7)) and imaginary (*Im*, Supplementary Equation (S8)), whose phases differ from each other by 90° (Supplementary Figure S1). To change the complex value into an absolute value, the following equation is applied:

$$|z| = \sqrt{Re^2 + Im^2} = \sqrt{\left(\gamma \cos \omega t \exp\left(-\frac{t}{T_2^*}\right)\right)^2 + \left(\gamma \sin \omega t \exp\left(-\frac{t}{T_2^*}\right)\right)^2} \quad (3)$$

For the matrix factorization method PSMF[109], positive-valued matrices are needed, and the original signal values must be converted to their logarithmic form for optimal analysis. To convert the absolute value in Equation (3) to a positive logarithmic form, the following Equation (4) (Supplementary Equation (S10)) is applied:

$$V = \log_{10}(|z| + 1) \quad (4)$$

Signal deconvolution can be then formulated as finding the factorization of the data matrix V (Supplementary Equations (S11) and (S12)):

$$V = W \cdot H + residuals = W_{signal} \cdot H_{signal} + W_{noise} \cdot H_{noise} + residuals \quad (5)$$

In this method using PSMF, we focus on sparse factorizations and on properly accounting for uncertainties while computing the factorization. Equation (5) estimates that the signal component ($W_{signal} \cdot H_{signal}$) decays exponentially with time, while the noise component ($W_{noise} \cdot H_{noise}$) is a random or flat value. To reconstruct the FIDs, the absolute value within each component is converted back to a complex value (Supplementary Equations (S13) and (S14)). The inverse STFT is computed by overlap-adding the inverse fast FT signals in each segment of the STFT spectrogram (Supplementary Equation (S15)).

To evaluate SNR, both noise-removed and noise-only FIDs are converted to signal and noise spectra, respectively, by applying standard FT. SNR is calculated as the ratio of the signal peak intensity to the noise value by using the method of Mnova (Supplementary Equation (S16))[110]. The noise value is calculated by using the standard deviation of the signals-free region (Supplementary Equation (S17)). Finally, the relative

SNR is the ratio of the SNR after denoising ($SNR_{denoised}$) to the original SNR ($SNR_{original}$), which is calculated as follows (Equation (6) and Supplementary Equation (S18)):

$$Relative\ SNR\ = \frac{SNR_{denoised}}{SNR_{original}} \quad (6)$$

Figure 1 shows an example of application of our signal deconvolution process to sucrose ^{1}H-NMR. STFT of the original FID adds a time axis to the frequency axis of the conventional FT spectrum (Figure 1a). The STFT spectrogram is three-dimensional, showing the frequency, time, and intensity of signal and noise. The matrix of the spectrogram was separated into signal and noise components based on the patterns of relaxation time using PSMF (Figure 1b). Each component was then converted into a signal FID and time-domain noise data by using inverse STFT (Figure 1c). Lastly, the time-region data were converted into the denoised spectrum and noise by using standard FT (Figure 1d). Regarding the noise reduction of the sucrose data, SNR of the denoised spectrum was improved about tenfold relative to the original data. In other words, for the sucrose sample, a 100-fold longer acquisition time would be required to obtain the same SNR without denoising. We compared signal and spectral quality between the original FT and noise reduction data (Supplementary Figure S2 and Table S1). There was almost no difference between them.

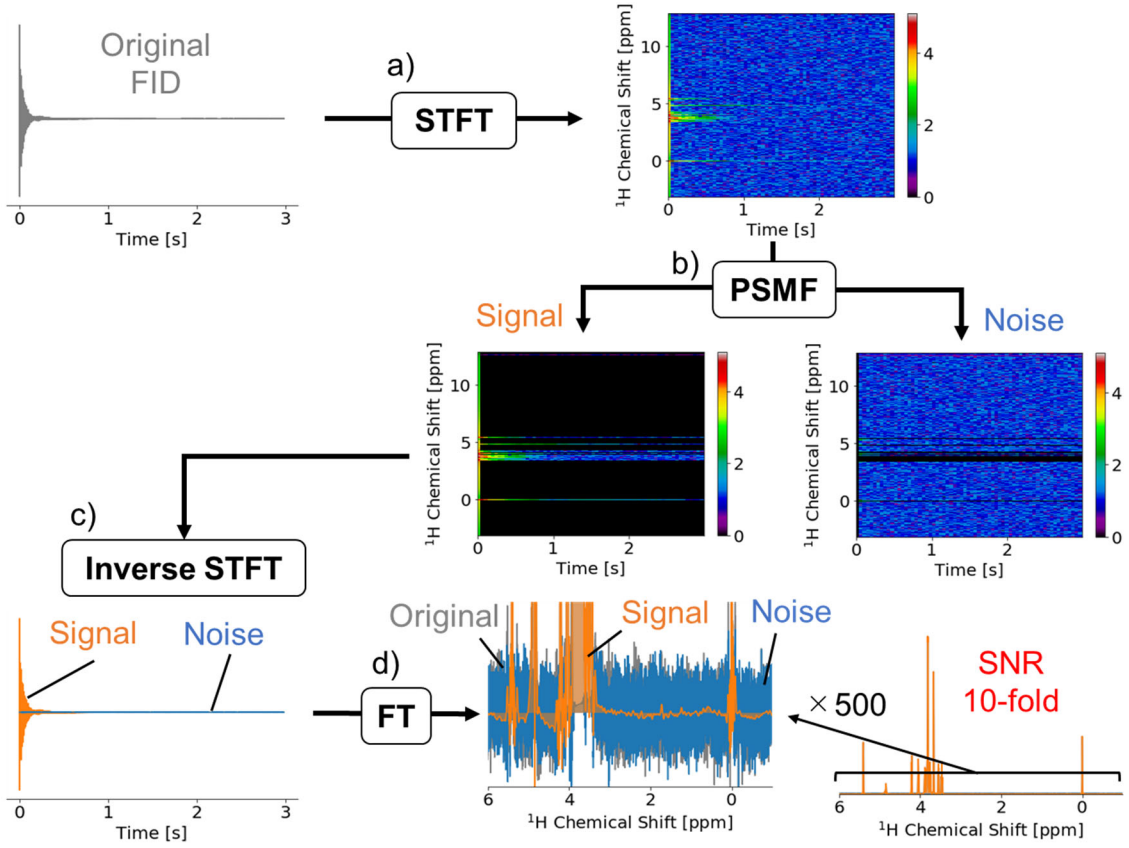


Figure 1. The free induction decay (FID) signal deconvolution method and its application to ^{1}H-NMR data for sucrose. (**a**) The spectrogram was obtained by applying short-time Fourier transform (STFT) to the original FID. (**b**) The matrix obtained after STFT was applied to probabilistic sparse matrix factorization (PSMF), which separated it into signal and noise components. (**c**) The signal and noise components were converted into a noise-removed FID signal (orange) and a time-domain noise signal

(blue) by using inverse short-time Fourier transform. (**d**) Finally, the noise-removed FID and the time-domain noise signal were converted to a frequency-domain spectrum by applying standard Fourier transform. As compared with the original FID, the signal-to-noise ratio of the denoised FID was improved about tenfold.

In STFT, the size of a window function $g(t-\tau)$ is important. We define the percentage of the time width as the percentage of the window size to FID length. After examining different percentages of the time widths, we found that signal components could be properly extracted in 1.5% and 3.1% (512 and 1024 points for 33280 points), but not 6.2% (2048 points for 33280 points) (Supplementary Figure S3). This is because the larger time width does not improve spectra since STFT becomes standard FT. Consequently, the percentage against the effective average region of FIDs is important for this method. Based on this result, the percentage of the time width was set to 3.1% for data analyzed in Figure 1. When using this method for data with short effective regions (fast relaxation systems such as solid-state NMR and quadrupole nucleus), data processing must be adjusted to maintain the shorter percentage of the time width. In addition, if an FID consists of a number of signals with differing T_2*, it will not be possible to choose an optimal filter for all lines simultaneously by applying commonly used apodization. The apodization such as exponential filtering decreases both signal and noise. In contrast, the method that we propose enables signal and noise to be extracted from an FID based on each pattern of T_2* relaxation time.

We compared the performance of PSMF with that of three other MF methods, namely standard nonnegative matrix factorization (NMF), sparse nonnegative matrix factorization (SNMF), and probabilistic nonnegative matrix factorization (PMF) (Figure 2). For PSMF, the noise region was successfully removed from the signal component (Figure 2a). For the other three methods, by contrast, the noise component remained in the signal component (Figure 2b–d). Regarding the PSMF time-varying coefficients, the signal component attenuated gradually over time, whereas the noise component attenuated sharply in the first segment and then became flat from the second segment (Supplementary Figure S4a). This observation suggests that part of the signal component may be included in the initial stage of the noise component. Therefore, for the optimal result in Figure 1, the initial value of the noise component is added as a signal component. The time-varying coefficients of the other three methods were characterized by containing mostly noisy components in the signal components, suggesting that the signal components were not properly extracted (Supplementary Figure S4b–d). The signal component is theoretically considered to be sparse data that comprise only specific frequency components. PSMF is a method that considers noise and uncertainty under the sparseness constraint, which suggests that it is suitable for removing noise from ^{1}H-NMR data. We also examined the effect of the number of components in PMSF on signal deconvolution, which showed that it was possible to properly extract signal components when there were two components (Supplementary Figure S5). When the number of components was increased, only noise components were separated more finely. Based on this result, the number of components was set to 2 in the signal deconvolution method for noise reduction. In the case of more complex data, such as the NMR signal of a mixture, it may be possible to apply the method to the characterization of multiple components by separating them with an arbitrary number of components.

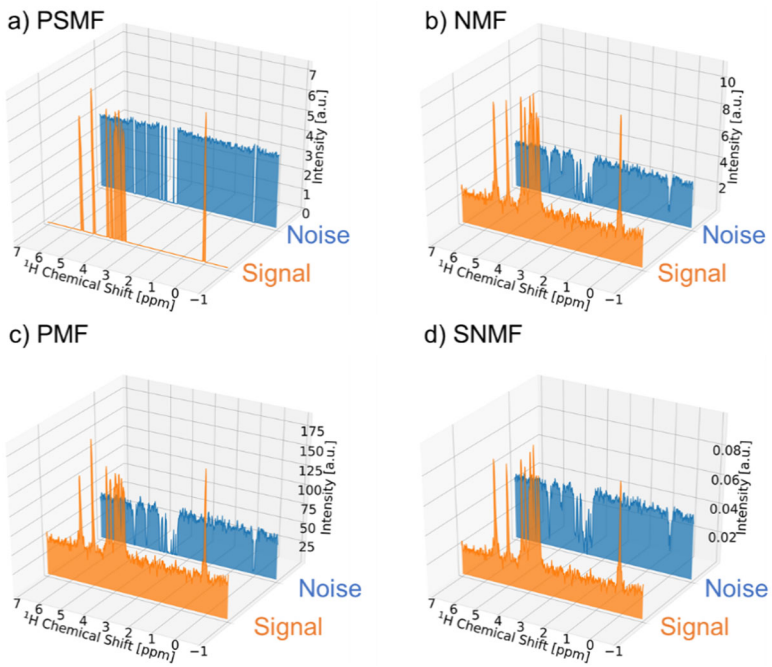


Figure 2. Comparison of four matrix factorization (MF) methods in signal deconvolution. Shown are spectral patterns of signal deconvolution for sucrose ^{1}H-NMR data using (**a**) PSMF, (**b**) NMF, (**c**) PMF, and (**d**) SNMF. The signal components are shown in orange and the noise components are shown in blue.

4.2.2 Noise reduction in NMR data measured by various pulse sequences

The improvement in the relative SNR achieved by the noise reduction method was investigated by using large-scale data measured by various pulse sequences (Figure 3). Here, we analyzed the following three pulse sequences, which are generally used depending on the target of analysis: CPMG, which detects small molecules with long T_2*, diffusion-edited, which detects proteins and lipids with relatively short T_2*, and WATERGATE, which detects both of these. For the analysis of extensive data, percentages of the time width to FID lengths were set to 6.3% for CPMG and WATERGATE, 12.5% for diffusion-edited (1024 points for 16384 and 8192 points), and the initial three values of the noise component were added as a signal component. For CPMG and WATERGATE, the improvement rate was 3.7-fold and 3.3-fold, respectively. On the other hand, it was only 2.2-fold for diffusion-edited NMR data (Figure 3a). As a result of comparing the relative SNRs of three typical pulse sequences for 10 representative samples, the data of diffusion-edited tended to be lower than those of CPMG and WATERGATE as in the case of large-scale data (Figure 3b, Supplementary Table S2) since the time width for diffusion-edited (12.5%) is higher than that of the other two pulse sequences (6.3%). The SNR of any NMR data set is related to the acquisition parameters (Supplementary Figures S6–8). In NMR data using CPMG and WATERGATE, the SNR is related to several acquisition parameters, such as receiver gain (RG), number of scans (NS), relaxation delay time (D1), spectral width (SW), and offset of the transmitter frequency (O1), whereas in diffusion-edited NMR, the SNR is

particularly related to the gradient pulse in the *z*-axis (GPZ). In diffusion-edited NMR, signals from small molecules with long T_2* relaxation times are suppressed. We therefore considered that, if the GPZ setting was insufficient, signals of small molecules would remain, resulting in a difference in relative SNR. As expressed, the peak SNR depends on T_2* because an FID with large T_2* yields a sharp line with higher SNR at the peak[100]. Thus, it seems likely that the diffusion-edited NMR data contain a lot of broad signals derived from macromolecules and lipids, resulting in less improvement as compared with CPMG and WATERGATE which have many sharp signals.

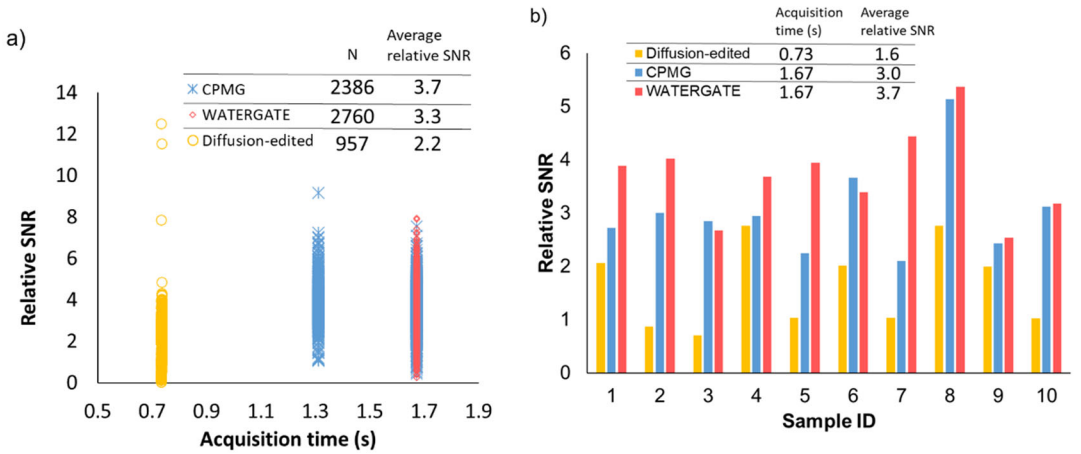


Figure 3. Relative SNR in data measured by three pulse sequences. (**a**) Shown is the relationship between the relative SNR after application of the noise reduction method to large-scale data measured by three pulse sequences: CPMG (blue), WATERGATE (red), and diffusion-edited (yellow), and its acquisition time. The upper part of the figure shows the number of spectra and the average relative SNR for each pulse sequence. (**b**) Comparison of the efficiency for improvement of the SNR measured by three pulse sequences: CPMG (blue), WATERGATE (red), and diffusion-edited (yellow), among NMR spectra derived from sample ID of 1 to 10. The acquisition time and the average relative SNR for each pulse sequence are shown in the upper part of the figure.

4.2.3 Application of signal deconvolution method in diffusion-edited NMR

We further examined the application of our signal deconvolution method to diffusion-edited NMR data. For the optimal analysis of these data, the percentage of the time width to FID length was set to 6.3% (512 points for 8192 points), and the initial value of the noise component was added as a signal component. The original FID was separated into three components, including noise and the long and short components of T_2* (Figure 4a,b). By extracting each component and performing standard FT, the SNR of the denoised spectrum was improved about threefold as compared with the original data. In addition, we obtained individual spectra for the short and long components of T_2* (Figure 4c,d). Thus, the diffusion-edited spectrum was separated into signals from macromolecules and lipids and small molecules by the length of the T_2* relaxation time. The composition of molecules in these signals is related to the GPZ value of the acquisition parameters (Supplementary Figures S8 and S9). We consider that insufficient GPZ is the main factor affecting the relative SNR of diffusion-edited NMR data because, if GPZ is insufficient, relatively more signals from small molecules are contained in the measured signals. Knowing this composition will help to evaluate the data quality of diffusion-edited NMR.

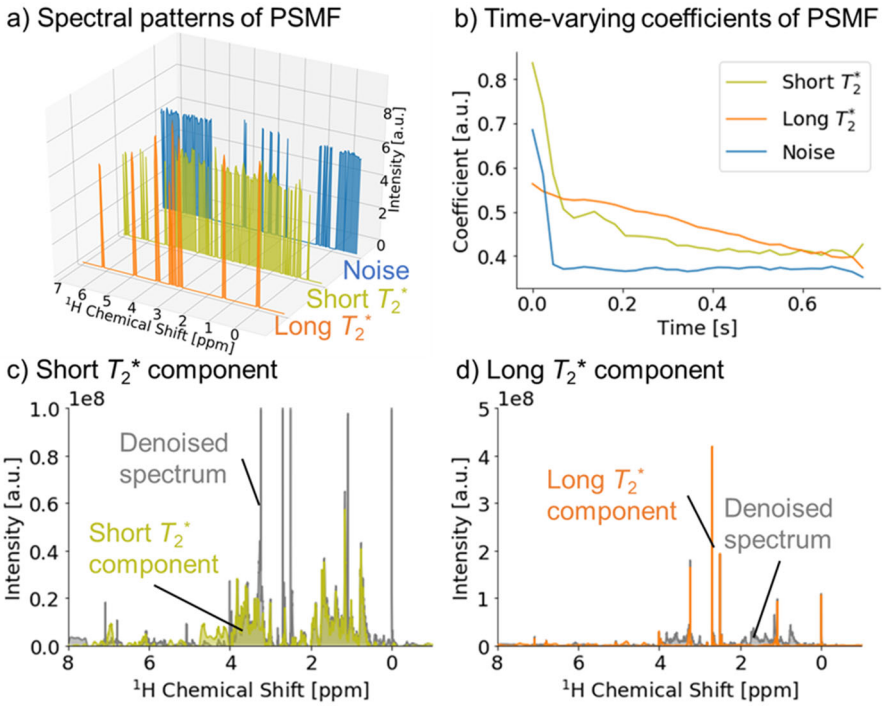


Figure 4. Application of the signal deconvolution method to diffusion-edited spectra. (**a**) Spectral patterns showing signals from small molecules (orange) and macromolecules and lipids (olive) separated by the length of the T_2* relaxation time, and noise (blue). (**b**) Time-varying coefficients of each component in MF. (**c**) Denoised spectrum (gray), and spectrum of the short T_2* component (olive). (**d**) Denoised spectrum (gray), and spectrum of the long T_2* component (orange).

4.2.4 Noise factor analysis in data measured by low- and high-field NMR at multiple institutions

To investigate the comprehensive relationship between noise and several acquisition parameters, we analyzed noise factors in data acquired by low- and high-field NMR at multiple institutions. We collected NMR data for four compounds (glucose, sucrose, citric acid, and lactic acid) measured by benchtop NMR (60 MHz) and high-field NMR (500, 600, and 700 MHz) from five institutions/data repositories (RIKEN, NUIS (Niigata University of International and Information Studies), BMRB[48], BML[49], and HMDB[47]) (Supplementary Table S3). The results of correlation analysis between noise and experimental parameters were first summarized as a heatmap (Supplementary Figure S10). With a specific focus on the experimental parameters that affect the SNR, we then derived a network of experimental factors affecting noise based on the correlation coefficients between SNR and experimental parameters (Figure 5). Here, in addition to the SNR calculated using Mnova, we calculated a theoretical SNR value (calcSNR) using a previously described SNR formula (Supplementary Equation (S19))[111] in order to obtain a theoretical SNR index based on acquisition parameters. Figure 5 shows that, based on the correlation between SNR and, for example, number of scans (NS) and signal intensity (e.g., standard, sample, and solvent), the integration of strong signals will

increase noise and reduce SNR. Therefore, the suppression of water signals and sample concentration will be important factors to obtain NMR data with a good SNR.

In situations where longer NMR measurements are needed owing to poor signals (e.g., for nuclei of low sensitivity and/or low natural abundance, and samples of low concentration), paying attention to the certain factors, as discussed here, may provide significant improvements in SNR[100], or even more marked savings in measurement time for a given SNR. For example, too long an acquisition time is not beneficial for SNR. An FID of the time constant T_2* gives, on Fourier transformation, a line width of $1/\pi T_2$* or approximately $1/3T_2$*. Thus, data acquisition beyond about $3T_2$* provides little gain in resolution, but causes a considerable deterioration in SNR. In addition, the spectral width may be set high enough to prevent aliasing of NMR signals. If not, there may be still other signals that fold, namely noise, meaning that the final SNR in the spectrum deteriorates.

Receiving efficiency (R) has been proposed as a way to characterize how efficiently the NMR signal can be observed after a unit transverse magnetization in a sample under optimal probe tuning and matching conditions[101]. In that study, the NMR signal amplitude was described as a function of the instrument constant, receiver gain, excitation angle θ, inhomogeneity factor $I(\theta)$, concentration of the observed nucleus, and sample volume. Modern NMR spectrometers require receivers to work within their linear ranges to maintain high-fidelity line shapes and peak integration[102]. The NMR receiver gain is a parameter that is often chosen to maximize SNR. For example, for optimal sensitivity, a dilute analyte needs to be observed with high NMR receiver gain, while the strong, interfering solvent signal must be suppressed[103]. In this case, the dependence of $I(\theta)$ on θ becomes more significant because homogeneity is typically lower for a cryoprobe than for its conventional counterpart[104], and failing to recognize the dependence of $I(\theta)$ on θ alone may potentially lead to errors in quantification as large as 5%. Other factors that we have discussed have less effect on SNR, but are significant in terms of line shape.

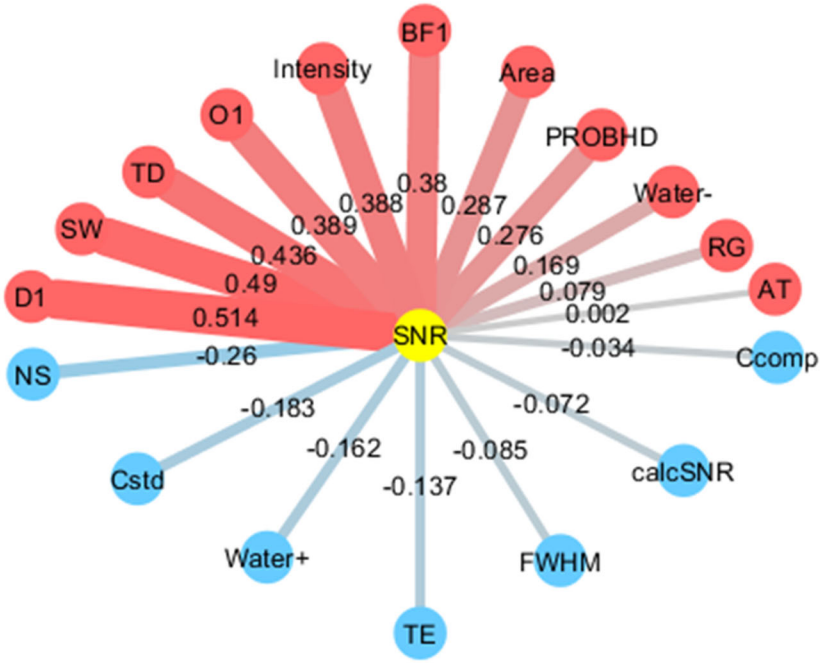


Figure 5. Analysis of experimental factors based on a correlation network of SNR and experimental parameters. The network diagram was drawn by setting positive correlations to red, negative correlations to blue, and the magnitude of the correlation coefficient to the edge thickness. Abbreviations: SNR, signal-to-noise ratio; calcSNR, calculated SNR; Cstd, concentration of standard compound; Ccomp, concentration of compound; Water+, positive intensity of water signal peak to standard peak; Water–, negative intensity of water signal peak to standard peak; Intensity, intensity of standard signal; FWHM, full width at half maximum; Area, area of standard signal; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; AT, acquisition time; TD, time-domain data size; O1, offset of transmitter frequency; TE, temperature; BF1, basic transmitter frequency for channel F1 in Hertz; PROBHD, if cryoprobe, value is 4, if not, value is 0.

3.4 Materials and Methods

4.3.1 Signal deconvolution method

The signal deconvolution method was developed in python 3, and built as a graphical user interface (GUI) tool using Tkinter. The tool is available on http://dmar.riken.jp/NMRinformatics/. The processing of NMR data was implemented by using the nmrglue[112] package in Python. PSMF[109], PMF[113], SNMF[114], and standard NMF[115] were calculated based on the NIMFA Python library for nonnegative matrix factorization[106].

4.3.2 Noise factor analysis method

The noise factor analysis consisted of four steps implemented in python 3, namely: (1) Collecting acquisition parameters of NMR data: FID and acquisition parameters were searched from the selected NMR data directory and written to CSV files.

(2) Calculating SNR: each FID was usually processed to an FT spectrum and denoised spectrum, and the SNR and its improvement ratio were calculated. In the noise factor analysis of data collected from multiple databases, SNR was calculated by using Mnova. (3) Calculating the correlation coefficient between SNR and each parameter by Pearson's correlation coefficient. (4) Visualizing experimental factors: the nodes, edges, and widths of networks based on the correlation coefficient were transformed in GML format by using the Networkx package in Python. Lastly, the network figure was drawn by using Cytoscape[116].

4.3.3 NMR data acquisition

Briefly, ^{1}H-NMR data were by recorded using an Avance II 700 Bruker spectrometer equipped with a 5-mm inverse CryoProbe operating at 700.153 MHz for ^{1}H. In the ^{1}H -NMR data, the number of data using CPMG pulse sequence was 2386, the number of data using WATERGATE pulse sequence was 2760, and the number of data in the 1D LED experiment using bipolar gradients (diffusion-edited) pulse sequence was 975 [33,117-119]. Regarding these large data sets, a summary of information on the sample and acquisition parameters (the sample title, solvent, acquisition time, acquisition point, and the original SNR) is available at http://dmar.riken.jp/NMRinformatics/. Data sets for comparing the relative SNRs of three typical pulse sequences for 10 representative samples are shown in Supplementary Table S2. To demonstrate the denoising method, data for sucrose and citric acid were acquired by using the presaturation (program name; "zgpr") pulse sequence. To demonstrate the method of separating signals in the diffusion-edited spectrum, ^{1}H-NMR data for fish muscle were measured by a diffusion-edited pulse sequence. Lastly, 48 sets of ^{1}H-NMR data (glucose, sucrose, citric acid, and lactic acid) were collected from the following five sites; RIKEN, NUIS, BMRB, BML, and HMDB. The data were measured with NMR spectrometers of 60, 500, 600, and 700 MHz manufactured by Bruker, Varian, and Nanalysis (Supplementary Table S3).

3.5 Conclusions

We have developed a measurement informatics tool for NMR signal deconvolution and noise factor analysis and used it to investigate the relationship between noise and acquisition parameters in accumulated NMR datasets. This method enables 1D-NMR spectra to be evaluated with a high SNR, and residual signals from small molecules to be removed from diffusion-edited spectra. This method can be adjustable to any T_2* length, recycle delay, sample molecular weight, or measurement temperature. The percentage of the time width against the effective average signal region of FIDs must be adjusted according to T_2* length. Therefore, when using this method for fast relaxation systems such as solid-state NMR and quadrupole nucleus, additional efforts are needed. In the case of 2D-NMR, it is necessary to use this method by splitting each t_1-dimensional FID and creating a series of sub-FIDs. Noise factor analysis of accumulated NMR datasets might facilitate the investigation of experimental factors related to a lower SNR. Therefore, these methods will help to determine optimal acquisition parameters, to

cleanse data, including data management and noise reduction in accumulated NMR datasets, and to promote data-driven studies of molecular complexity using NMR.

Part IV

Material development approach using solid-state NMR

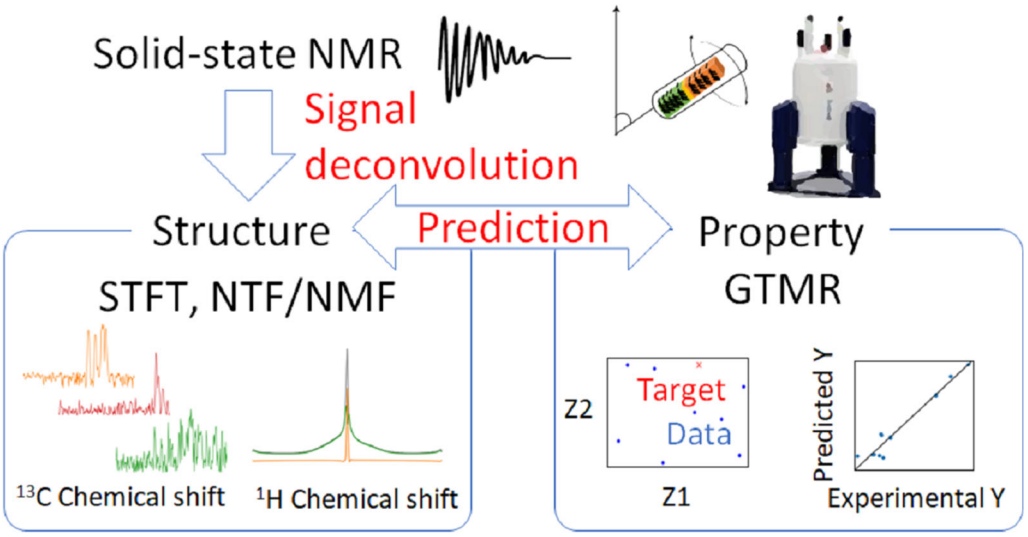
Chapter 4

Signal Deconvolution and Generative Topographic Mapping Regression for Solid-State NMR of Multi-Component Materials

4.1 Abstruct

Solid-state nuclear magnetic resonance (ssNMR) spectroscopy provides information on native structures and the dynamics for predicting and designing the physical properties of multi-component solid materials. However, such an analysis is difficult because of the broad and overlapping spectra of these materials. Therefore, signal deconvolution and prediction are great challenges for their ssNMR analysis. We examined signal deconvolution methods using a short-time Fourier transform (STFT) and a non-negative tensor/matrix factorization (NTF, NMF), and methods for predicting NMR signals and physical properties using generative topographic mapping regression (GTMR). We demonstrated the applications for samples involved in cellulose degradation, plastics, and microalgae such as *Euglena gracilis*. During cellulose degradation, ^{13}C cross-polarization (CP)–magic angle spinning spectra were separated into signals of cellulose, proteins, and lipids by STFT and NTF. GTMR accurately predicted cellulose degradation for catabolic products such as acetate and CO_2. Using these methods, the ^{1}H anisotropic spectrum of poly-ε-caprolactone was separated into the signals of crystalline and amorphous solids. Forward prediction and inverse prediction of GTMR were used to compute STFT-processed NMR signals from the physical properties of polylactic acid. These signal deconvolution and prediction methods for ssNMR spectra of

macromolecules and lipids can resolve the problem of overlapping spectra and support the characterization and design of materials.



Graphical abstract

4.2 Introduction

Recently, research for a low-carbon society has gained importance from the viewpoints of global challenges such as the marine pollution of marine plastics, waste disposal, and global warming[120]. Microbial products and plant biomass as alternatives to petroleum resources can be used to produce macromolecular materials such as plastics and feedstock[121]. Polymers such as polylactic acid (PLA)[122], poly-ε-caprolactone (PCL)[123], and cellulose[124-131] are multiple domain/component systems and are often employed as high-performance materials with various properties. Microbial and plant biomass should be analyzed as a biochemical system composed of multiple components containing macromolecules and lipids with multiple domains. Solid-state nuclear magnetic resonance (ssNMR) spectroscopy is a powerful tool for characterizing the native structure, components, and dynamics of solid-state samples at the atomic level. It is being increasingly applied in material/life sciences[5,132]. Therefore, an advanced ssNMR analytical approach must be developed for products such as microbial products, plant biomass, and plastics.

Various techniques that use high-field magnets, cryogenic detection systems, indirect detection[133], nonuniform sampling[134], and dynamic nuclear polarization methods[135,136] have been developed for realizing increased sensitivity. From the aspect

of NMR measurement, various solid-state NMR methods have been used. Typical methods are cross-polarization (CP)–magic-angle spinning (MAS) methods, static multiple-quantum (MQ) NMR, static ^{1}H NMR[137], direct polarization (DP), high-resolution (HR)-MAS[29,45,66], magic-and-polarization echo (MAPE) filtering[138], double-quantum (DQ) filtering[139], and combined rotation and multiple-pulse techniques (CRAMPS)[140]. MAS probes are capable of spinning frequencies much greater than 100 kHz[141]. Other advanced techniques are spin diffusion measurements[11], pulsed field gradient (PFG) NMR, diffusion-ordered spectroscopy (DOSY), and time-domain NMR/relaxometry[142]. In addition, multi-dimensional NMR was applied for separating overlapping spectra; examples of such techniques are wide-line separation (WISE) and heteronuclear correlation (HETCOR)[143,144], three-dimensional (3D) dipolar-assisted rotational resonance, double-cross-polarization ^{1}H-^{13}C correlation spectroscopy, and ^{1}H–^{13}C solid-state heteronuclear single-quantum correlation spectroscopy[66].

In the characterization of solid-state samples with crystal, interphase, and amorphous domains, the anisotropy detected by static measurement is useful, but its analysis is difficult because the spectra are broad and overlapping[145]. Therefore, the application of signal deconvolution to measure solid-state NMR data is an important challenge to extract hidden information in the NMR spectra of samples with multiple phases and components. Several methods for spectral separation [94], apodization, zero filling, linear prediction, fitting and numerical simulation[146], such as covariance analysis[147], SIMPSON[148], SPINEVOLUTION[149], dmfit[150], EASY-GOING deconvolution[151], INFOS[152], Fityk[153], ssNake[154], the noise reduction method based on principal component analysis[95], and the signal deconvolution method that combines short-time Fourier transform (STFT, a time–frequency analytical method), and probabilistic sparse matrix factorization (PSMF which is one of the non-negative matrix factorizations)[155] were developed as computational approaches to measured data.

In this study, we propose signal deconvolution methods using STFT and non-negative tensor/matrix factorization (NTF, NMF) optimized to characterizing the solid-state NMR spectra of samples with multiple domains and components such as cellulose, plastics, and *Euglena gracilis*. Using generative topographic mapping regression (GTMR, the regression method using GTM)[156], we mutually predicted higher-order structure descriptors of STFT-processed NMR signals (STFT–NMR signals) and physical properties of the material. To the best of our knowledge, this is the first reported application on the prediction of NMR signals from the thermal properties of plastics using GTMR.

4.3 Results and Discussion

5.2.1 Signal deconvolution and prediction for solid-state NMR of multi-component materials

In this study, from a practical point of view, we focused on a signal deconvolution method for one-dimensional (1D) ssNMR data suitable for high-throughput multi-sample measurement. In particular, static ^{1}H anisotropic spectra can be used as an index of the motility of higher-order structures, but these spectra are broad and show overlapping. Even extremely sharp spectra such as ^{13}C CP-MAS show overlaps, especially in the case of signals with different mobility derived from the same atom. Therefore, those data must be separate signals. In principle, the exponential decay constant of the free induction decay (FID) obtained by applying a 90° pulse to create transverse magnetization is the T_2 relaxation time. In reality, however, because of the effect of magnetic field inhomogeneity, the decay constant of the FID is defined as T_2^*, an instrument-dependent parameter, rather than T_2. In this paper, we report a signal deconvolution method to separate the broadening spectra derived from cellulose and plastics with multiple phases and components based on the T_2^* relaxation pattern. The short-time Fourier transform (STFT) method is used to convert an FID into frequency domain data at short time intervals to generate a matrix of time and frequency axes (Figure 1a). As algorithms of factorization, in addition to the traditional NMF for analysis of the two-dimensional (2D) dataset, we investigated the application of NTF (non-negative Tucker decomposition (NTD)[157] and non-negative canonical polyadic decomposition (NCPD)[158,159]), which is a factorization algorithm useful for the analysis of the 3D dataset of multiple samples and parameters. By applying NTF/NMF (Figure S1) to the dataset, the signal components were separated based on the T_2^* relaxation pattern of the components indicated in the multi-phase and multi-component spectra (Figure 1b,c). Furthermore, the high-order structure of materials exerts a significant influence on their macroscopic properties[11]. Traditional design approaches for materials are experimentally driven and trial-and-error are facing significant challenges due to the vast design space of materials. In addition, computational technologies such as density functional theory (DFT)[160] and molecular dynamics (MD)[126] are usually computationally expensive and are difficult to calculate molecular structures from material properties. To address these problems, machine-learning-assisted materials design is emerging as a promising tool for successful breakthroughs in many areas of science[12]. In addition, NMR measurement, especially a low magnetic field NMR, is a method for routine material evaluations, which produce a lot of NMR datasets[94]. Against this background, in the cycle of developing materials using NMR and other measurements, the prediction of the NMR signal using the accumulated data is necessary to find a structure with the desired properties. In this study, prediction of the NMR data and sample properties was calculated using GTMR (Figure 1d,e and Figure S2)[156]. For cellulose degradation samples, our previous study reported that solution ^{1}H and ^{13}C NMR data were used for evaluating the concentration of catabolic products. In this study, we examined the use of pseudodata as a method of predicting data without experiments. Pseudodata are a dataset with the same distribution as the original dataset generated using Gaussian mixture models (GMM) (Figure S3)[161]. Randomly generating data based on means and covariances using GMM produces new pseudodata. By

performing GTMR calculation from these pseudodata as input data, a spectrum as output can be predicted without preparing new materials. The STFT–NMR signals were predicted as a higher-order structure descriptor and were transformed to predicted NMR properties. This method can be applied to various sample systems for pursuing structure–property correlation. In this study, we demonstrate the application of cellulose degradation and plastic for evaluating our method. Here, in cellulose degradation, the word "higher-order structure" means the crystalline and amorphous structure of cellulose, and the word "property" means the quantity of catabolic products. In addition, with plastics such as PCL, it is difficult to design those having both high degradability and toughness. In the PCL, multiple domain structures with different degrees of entanglement of molecular chains are referred to as "higher-order structures", and thermal and mechanical properties are referred to as "property". This analytical flow is useful for the research and development of macromolecules and related products.

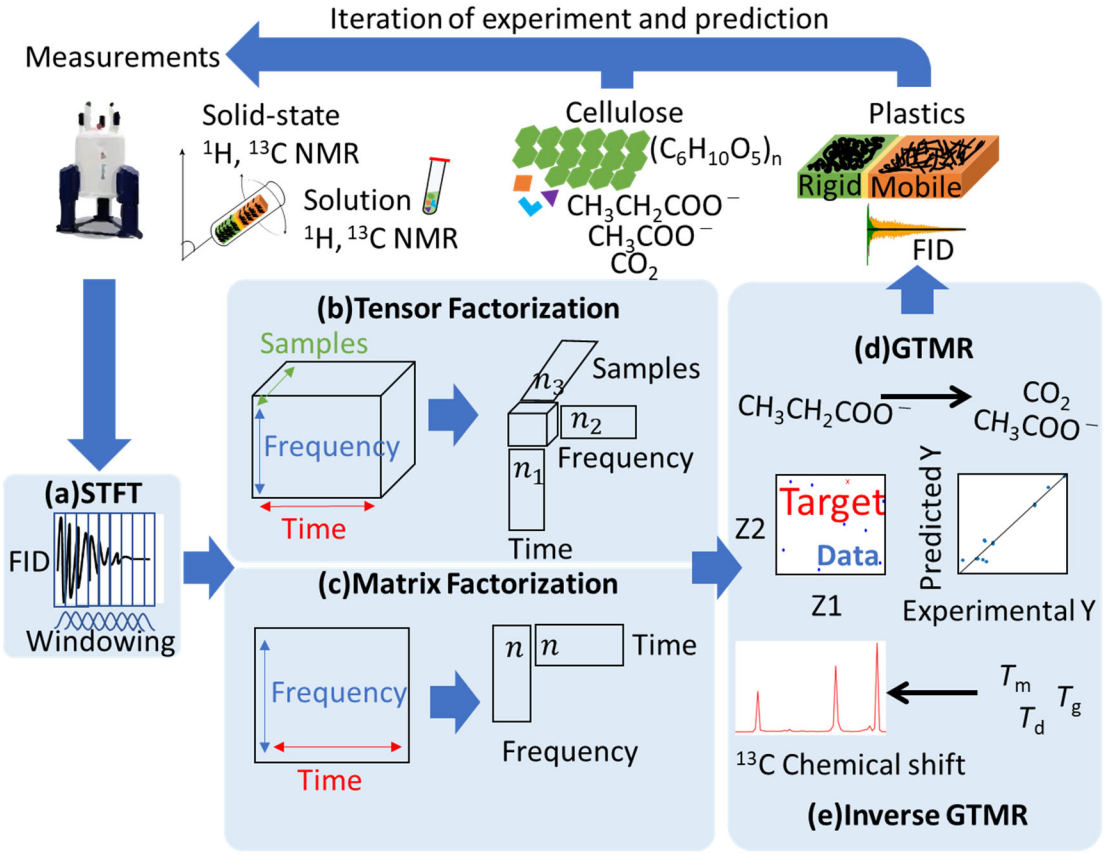


Figure 1. Concept diagram of a material development cycle based on signal deconvolution and prediction for the solid-state nuclear magnetic resonance (ssNMR) of multi-component materials. (**a**) Free induction decay (FID) is transformed into a dataset with time and frequency axes by short-time Fourier transform (STFT). (**b**) In the case of a three-dimensional dataset such as one with multiple samples and conditions, the FID is separated into each component based on the factors of time, frequency, and samples (or condition) by tensor factorization. (**c**) In the case of two-dimensional datasets such as a matrix with time and frequency axes, the FID is separated into each component based on factors of time and frequency by matrix factorization. (**d**) The generative topographic mapping regression (GTMR) accurately predicted the cellulose degradation process shown by catabolic products such as acetate and CO_2. (**e**) Forward prediction and inverse prediction of GTMR

were used to compute the STFT-processed NMR (STFT–NMR) signals from the physical properties of the plastics. This approach is an iterative procedure to achieve convergence between experimental and predicted spectra.

5.2.2 Non-negative Tucker decomposition to ^{13}C CP-MAS in cellulose degradation process

Solid and solution NMR methods can monitor higher-order structural changes and catabolic products during the degradation of cellulose by microorganisms[129,131]. The dataset used in Figure 2 is a time-dependent dataset of ^{13}C solid-state CP-MAS signals of the cellulose degradation process and also contains signals of catabolic products (proteins and lipids). The ^{13}C ssNMR spectra detect cellulose, proteins and lipids. This dataset is a set of data with frequency and intensity in 16 time points from 0 to 120 h (Figure 2a). This dataset was processed by STFT (Figure S4). We demonstrated the application of NTD (Figures 1b and 2b), which is one of the tensor factorizations for multi-sample data. By separating the spectrum into four components, it was possible to visualize the spectral patterns (Figure 2c–f), time change of each component (Figure 2g), and the composition (Figure 2h). The word "Time change" in Figure 2g means the change in acquisition time of the separated signal components. In addition, the word "Composition" in Figure 2h means the change in the 16 samples from 0 to 120 h of ^{13}C CP-MAS NMR spectra. As a result, the four signals (the cellulose, proteins, and lipids-like signals) were clearly separated as intense signals, while the noise was relatively low. In the calculation scheme of NTD, the convergence tolerance of calculation error was less than 0.001. The cellulose-like spectrum had a short relaxation time (Figure 2c,g (orange)), the protein-like spectrum had a long relaxation time (Figure 2d,g (green)), and the lipid-like spectrum had the longest relaxation time (Figure 2e,g (red)); the noise did not change. It was possible to evaluate the concentration of each component among samples (Figure 2h). As a result of separating the spectrum of the cellulose C4 region (Figure S5a) into six components, it was possible to visualize the spectral patterns (Figure S5b), time change of each component (Figure S5c), and the composition in each sample (Figure S5d). So far, tensor factorizations have been reported for the application of NCPD to solution NMR of carbohydrate mixtures[158] and high-dimensional NMR of protein structures[159]. As a result of separating the spectrum into four components using NCPD, it was not as good as NTD because of unclear spectral patterns for assigning compounds (Figure S6). NCPD is different from the algorithm of NTD used in this work. NTD separates the tensor into a small core tensor and factor matrices. NCPD separates the tensor into factor matrices without a core tensor. This study shows that the NTD is also effective for analyzing time-series ssNMR data such as those of the cellulose degradation process.

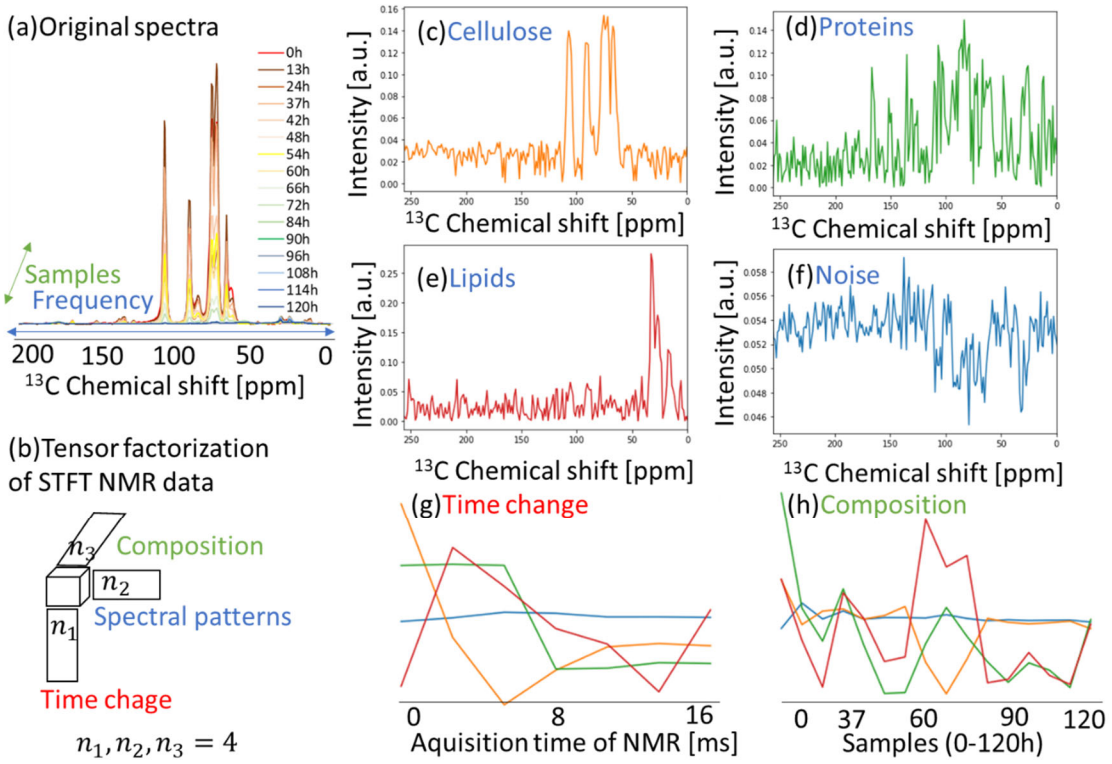


Figure 2. Application of non-negative Tucker decomposition (NTD) to ^{13}C cross-polarization–magic-angle spinning (CP-MAS) in the cellulose degradation process. (**a**) Original spectra of ^{13}C CP-MAS in cellulose degradation process. (**b**) Tensor factorization of STFT–NMR signals. (**c**–**f**) Spectral patterns (cellulose, lipids, proteins, and noise) when signals were separated into four components. (**g**) Time change of separated components. (**h**) Composition of separated components.

5.2.3 Non-negative matrix factorization to static ^{1}H ssNMR in PCL and *E. gracilis* Samples

PCL has a high-order structure of mobile, rigid, and interphase[142,146]. Evaluating the structure, motility, and proportion of multiple domains is important for material development including such as the optimization of physical properties. In the development of plastics especially, evaluation of higher-order structures is useful for the static ^{1}H anisotropic spectrum in solid states. From the aspect of the pulse program, by using a DQ filter or MAPE filter, components with different motilities can be extracted. In this study, we demonstrated the application of NMF to a 2D dataset created from the single data of PCL using STFT. Unlike NTF for a 3D dataset mentioned above, NMF is a method for a 2D dataset. NMF discovers hidden patterns in the axes of both time and frequency created by STFT, which is able to separate NMR signals to multiple components with different T_2^*. It was shown that by using NMF, rigid and mobile phases can be extracted from a broad static ^{1}H anisotropic spectrum of PCL as the components related to different physical properties (Figures 1c and 3). We resolved the linear macromolecular structure as a mobile domain and the branched macromolecular structure due to strong anisotropic ^{1}H-^{1}H dipolar coupling as a rigid domain in solid material such

as PCL. Furthermore, we demonstrated this method for ^{1}H, ^{13}C, ^{15}N and ^{31}P spectra of microalgae such as *E. gracilis* in a multi-component system (Figure S7). ^{1}H high-speed magic-angle spinning (MAS) spectrum was separated into signals of amide protons and fatty acids in lipids, and the ^{13}C CP-MAS spectrum was separated into signals of paramylon, lipids, and proteins. To overcome the limitation of sensitivity in NMR, various techniques were developed using high-field magnets, cryogenic detection systems, indirect detection[133], nonuniform sampling[134], and dynamic nuclear polarization methods[135]. We previously demonstrated that the STFT can be used for signal improvement of the solution diffusion-edited NMR spectra, including broad signals and sharp signals[155]; in this study, we demonstrated signal deconvolution using the STFT in the solid-state NMR. When using this method for NMR data with low digital resolution such as solid-state NMR and quadrupole nucleus, this signal deconvolution method needs additional efforts. We demonstrated some interpolation methods for increasing data points (Figure S8). The Fourier interpolation method provides an interpolated spectrum without artifact signals. Spectra interpolated by other methods have artifacts in the extended region.

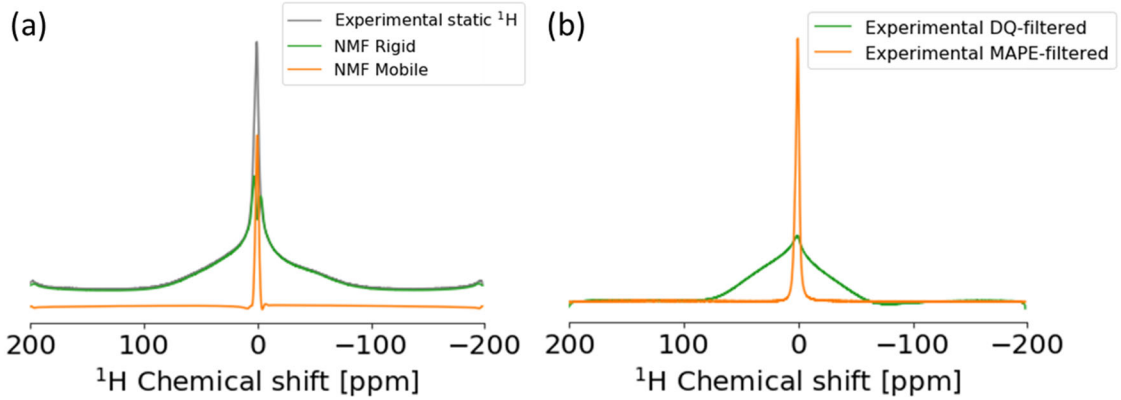


Figure 3. Application of non-negative matrix factorization (NMF) to static ^{1}H solid-state NMR of poly-ε-caprolactone (PCL). (**a**) Experimental anisotropic spectrum (gray) and spectra of rigid (green) and mobile (orange) components separated by NMF. (**b**) Experimental spectra of double-quantum (DQ) filtered ssNMR (green) and magic-and-polarization echo (MAPE) filtered ssNMR (orange).

5.2.4 Prediction of concentration of products in the cellulose degradation process

Thus far, GTM has been applied to characterize NMR data[162]. Recently, computational approaches for predicting NMR signals[160], chemical structures[163], and physical properties[164-168] were developed. Chemical shifts of NMR are rich in chemical information and enable encoding the structural features of the molecules contributing to their physical/chemical/biological properties. Thus, it has potential for use as a descriptor in quantitative structure–activity/property relationship (QSAR/QSPR) modeling studies[169]. GTMR was applied for analyzing these studies[156]. Therefore, the prediction of NMR signals is important for developing materials. This study is the first application of GTMR for the prediction of NMR signals (Figure 1d). In the degradation of cellulose, cellulose is metabolized into microbial cell components such as proteins and lipids, and then catabolized into short-chain fatty acids. In Figure 2, cellulose, proteins

and lipids were detected using the solid ^{13}C spectrum. In addition, to track the process of material degradation, solution NMR spectra were used to detect small molecules such as propionate and acetate. Therefore, the catabolic products were captured by solution NMR (the final products are CO_2 and CH_4 with one carbon atom (Figure S9)). During GTMR, multi-dimensional and multi-component data (in this case, CP-MAS data of macromolecules and lipids and small-molecule solution NMR data) can be mapped into the reduced dimensional space (Figure 4a,b left). When cellulose is finally catabolized to CO_2 by the catabolism of microorganisms, it is metabolized into acetate with two carbon atoms and CO_2 with one carbon atom via propionate with three carbon atoms. When the signal intensity of propionate is used as the input data of GTMR, it is possible to predict both the properties (scaled signal intensities in these results) of acetate (Figure 4a right; $R^2 = 0.976$) with the two carbon in the previous stage of the final product and CO_2 (Figure 4b right; $R^2 = 0.967$) with one carbon in the final product. GTMR thus provides information about the predicted NMR scaled signals of products in cellulose degradation. This information is important for monitoring the degradation process due to a key in compound production using cellulose.

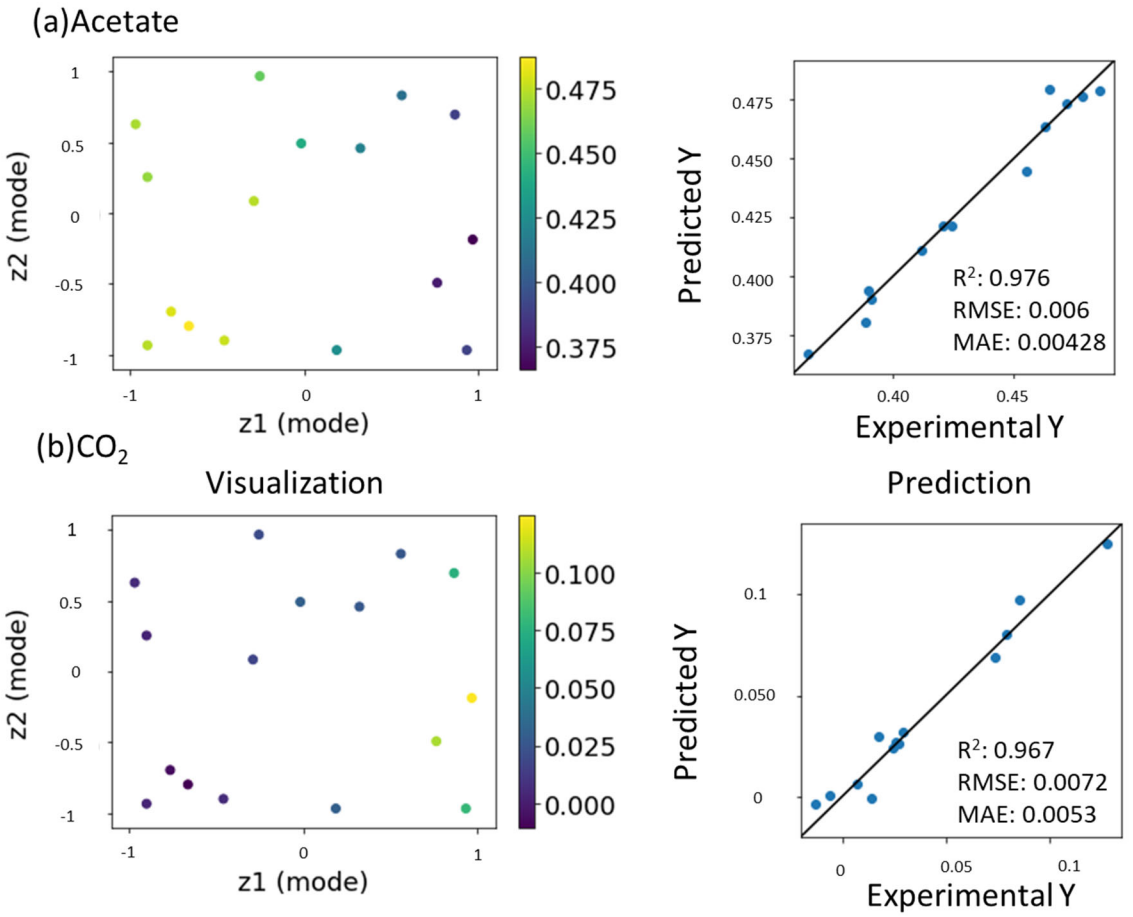


Figure 4. Application of GTMR to NMR data in the cellulose degradation process. (**a**) Visualization and prediction of the concentration of acetate. (**b**) Visualization and prediction of the concentration of CO_2.

5.2.5 Prediction of NMR signals from thermal properties in plastics

This study is the first application to predict NMR signals from the thermal properties of plastics using GTMR. The design method for higher-order structures of plastics should control the glass transition, melting, and degradation temperature (T_g, T_m, and T_d) as thermal properties. The GTMR was first applied for the inverse analysis of the CP-MAS spectra (Figure S10) from the thermal properties (Figure S11) of PLA in the solid state (Figure 1e). Therefore, T_g (Figure 5a), T_m (Figure 5b), and T_d (Figure 5c) were mapped into a reduced 2D space. We focused on the prediction of the intended thermal property (Figure 5d; red cross) using the three GTMR maps (T_g, T_m, and T_d). Hence, the STFT–NMR signals, i.e., the predicted spectrum, corresponded to the red cross and were predicted as higher-order structure descriptors (Figure 5e). Moreover, as a result of predicting the thermal properties from pseudo-CP-MAS spectra of PCL using GMM, it was possible to predict thermal properties (Figure S12).

Recently, the materials informatics (MI) approach was considered for material design[170] because the intended physicochemical property is really hard to identify in the material development process. Therefore, the MI approach uses “big-data” such as deposited database, as well as monitoring and analyzing higher-order structural data during the materials production process[171,172]. When developing a material with the desired physical properties, the molding conditions of the material with the predicted structure play an important role.

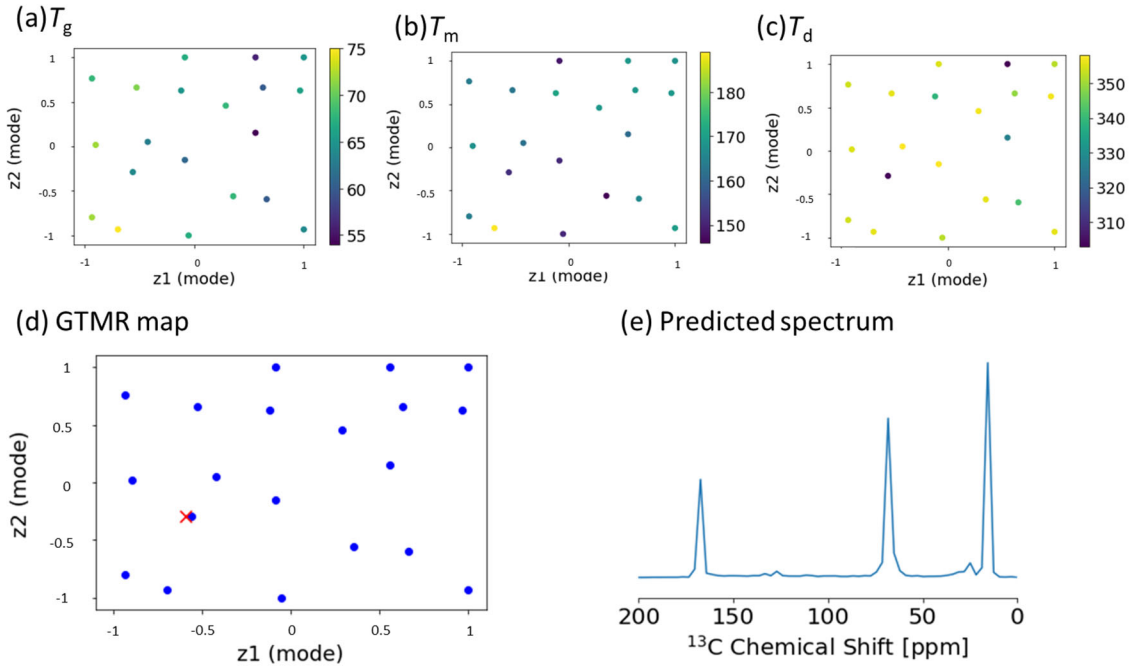


Figure 5. Application of GTMR for predicting NMR data from thermal properties in PLA. (a–c) Tg, Tm, and Td in data map. (d) Coordinates corresponding to the target thermal properties in data map. (e) Predicted 13C CP-MAS spectrum using GTMR.

4.4 Materials and Methods

4.4.1 NMR analysis

The ssNMR data were acquired using an Avance III HD-500 spectrometer (Bruker Corp., Billerica, MA, USA) equipped with a double-resonance 4.0 mm MAS probe. The solution NMR data were acquired using an Avance III HD-700 spectrometer (Bruker Corp., Billerica, MA, USA). The ^{1}H and ^{13}C CP-MAS spectra and solution ^{1}H and ^{13}C NMR spectra of cellulose previously reported by Yamazawa et al. were used[129]. The multiple phases polymer such as PCL, were measured using static, MAPE-filtered and DQ-filtered ssNMR. The ^{1}H, ^{13}C, ^{15}N, and ^{31}P spectra of *E. gracilis* cell previously reported by Komatsu et al. were used[66].

4.4.2 Thermal analysis of plastics

Thermogravimetry (TG) and differential thermal analysis (DTA) measurements were conducted using an EXSTAR TG/DTA 6300 (SII NanoTechnology Inc., Tokyo, Japan) instrument[26,143]. Approximately 10 mg of samples was individually vaporized at 5 °C/min from 40 to 500 °C in a nitrogen atmosphere. The T_m and T_d were determined as the endothermic peak in DTA curves and the peak of weight loss in Derivative Thermogravimetry (DTG) curves. Differential scanning calorimetry (DSC) was conducted using a DSC3500A (NETZSCH Geratebau GmbH, Selb, Germany)[173]. Approximately 1.5 mg of samples was individually measured at the following steps at 10 °C/min from 25 to −30 °C, at 10 °C/min from −30 to 200 °C, and at 20 °C/min from 200 to 25 °C in a nitrogen atmosphere. The T_g was determined as an endothermic peak during heating.

4.4.3 Signal deconvolution methods

The signal deconvolution method was developed in Python 3. The processing of NMR data was implemented by using the nmrglue[112] package in Python. Tensor factorization methods of NTD and NCPD were calculated using TensorLy Python library for tensor methods[157], and NMF was calculated based on the NIMFA Python library for non-negative matrix factorization[106]. NMR data with interpolated data points were created using “signal” and “interpolate” in “scipy”.

4.4.4 Prediction methods

Predictions of NMR signals and properties were calculated using GTMR[156]. In the analysis of cellulose degradation, a regression model was created using STFT–NMR signals, and product peak intensities were determined by solution NMR. As input data to analyze in GTMR, pseudodata were generated using GMM[161]. In the case of GTMR in the data of cellulose degradation process, the peak of propionate as input data was used, and the peaks of CO_2 and acetate were predicted as the concentration of production. For plastics analysis, a regression model was created using the STFT–NMR signals and thermal properties. In the case of inverse GTMR, the desired thermal properties were used as input data, and NMR signals were predicted as the higher-order structure descriptors.

4.5 Conclusions

We have developed a solid-state NMR signal deconvolution method using STFT and NTF/NMF, and a prediction method using GTMR. These methods enable 1D solid-state NMR spectra to provide separate signals of multiple phases and components from solid-state NMR spectra. Further, macromolecular samples were characterized, and higher-order structures and thermal properties were predicted. As a new alternative to applying the decoupling to remove anisotropy as unnecessary information in the measurement of ssNMR with a broad line width, signal separation by computational science methods will expand the applicability of low-field ^{1}H ssNMR and anisotropic NMR. In the case of NMR data with low digital resolution such as the solid-state NMR and quadrupole nucleus the number of data points can be increased by applying interpolation. In the case of 2D-NMR, it is necessary to use this method by splitting each t1-dimensional FID and creating a series of sub-FIDs. Therefore, these methods will promote data-driven research and development in fields such as machine learning and simulation using ssNMR on macromolecular complexity in materials and foods.

Part V

General Discussion

Chapter 5

Summary and Prospects

5.1 Summary

Currently, data-driven science is drawing attention. Therefore, I focused on the quality of data in data-driven science (Figure 1). Although NMR analysis of molecular complex systems can be applied to various fields, its use has been limited in terms of sensitivity and resolution. In this study, I tried to overcome these problems in NMR measurement by utilizing informatics.

<Originality of this work>

Chapter 2:

- Signal assignment by combined use of peak enhancement and matrix factorization
- NNSC as a new NMR signal separation method
- Database integration of ^{1}H-^{13}C and ^{1}H-*J* correlation, ^{13}C CP-MAS spectra
- ^{13}C CP-MAS spectrum assignment tool for macromolecules and lipids
- Signal assignment tool for two spectra of ^{1}H-^{13}C and ^{1}H-*J* correlations

Chapter 3:

- Signal deconvolution method that combines STFT and PSMF
- Visualization of quality and factors focusing on parameters and noise in NMR data

Chapter 4:

- Signal deconvolution method using STFT and NTF
- Application of GTMR in NMR

<Conceptual progress>

- Data management

  Identification of factors that affect individual data quality is possible by focusing on measurement parameters and noise.

- Signal assignment of low-resolution NMR spectra

  Assignment of overlapping peaks is possible by utilizing peak enhancement, spectral separation, signal deconvolution, and assignment tools.

- Signal deconvolution based on STFT and matrix and tensor factorization

  Signal deconvolution of measurement data by computational scientific methods is possible as an alternative to physical separation and various pulse sequence such as decoupling and diffusion-edited, which is applied to remove unnecessary signals before or during measurement.

- Prediction method for NMR signals and physical properties of materials

  Efficient material development is possible by computationally predicting NMR signals of materials with the desired properties (metabolic products and physical properties).

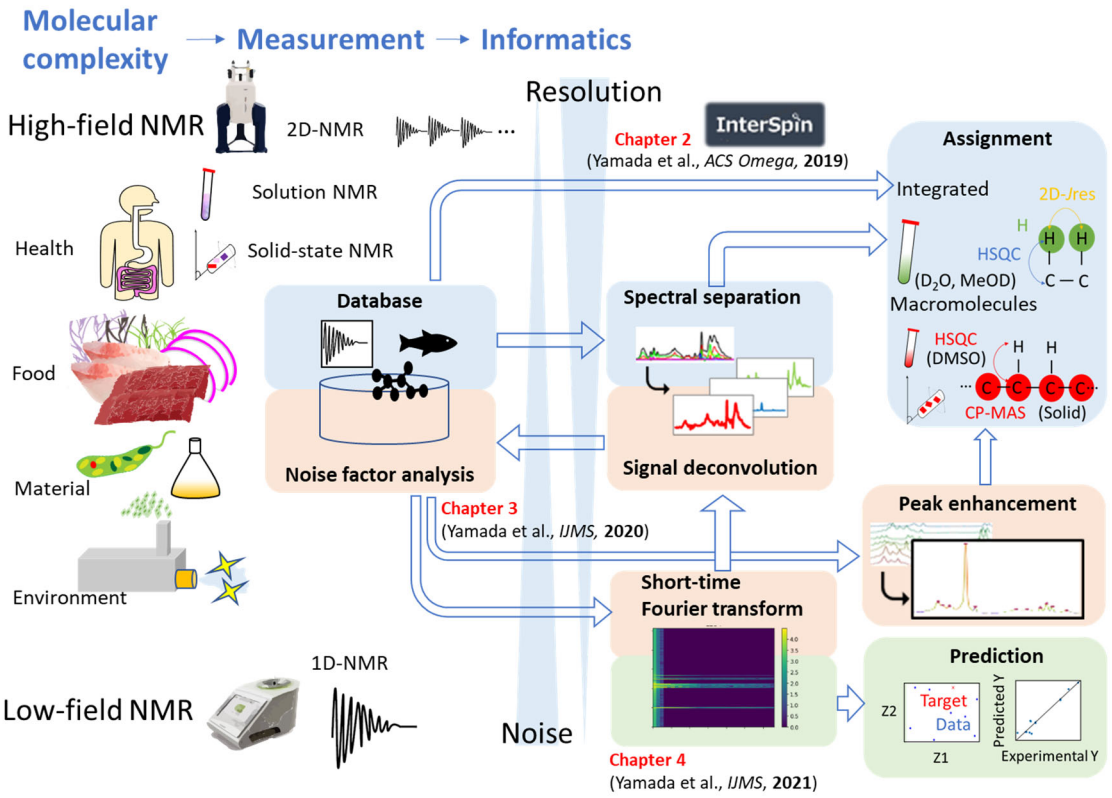


Figure 1. Summary of measurement informatics approaches in this doctoral research. This figure shows the position of the contents of chapters 2-4 in the measurement informatics flow of NMR.

5.2 Prospects

NMR data is becoming more and more important in social innovation. Large-scale data may be accumulated from high-field or low-field NMR (Figure 2). The data circulation system that organizes and effectively utilizes such data is immature. It is expected to support research by data-driven science such as simulation and machine learning by effectively connecting data and recommending data that is needed, and by combining noise reduction that enhances the quality of data. It is expected to develop a research approach that utilizes not only individual research but also research accumulated all over the world.

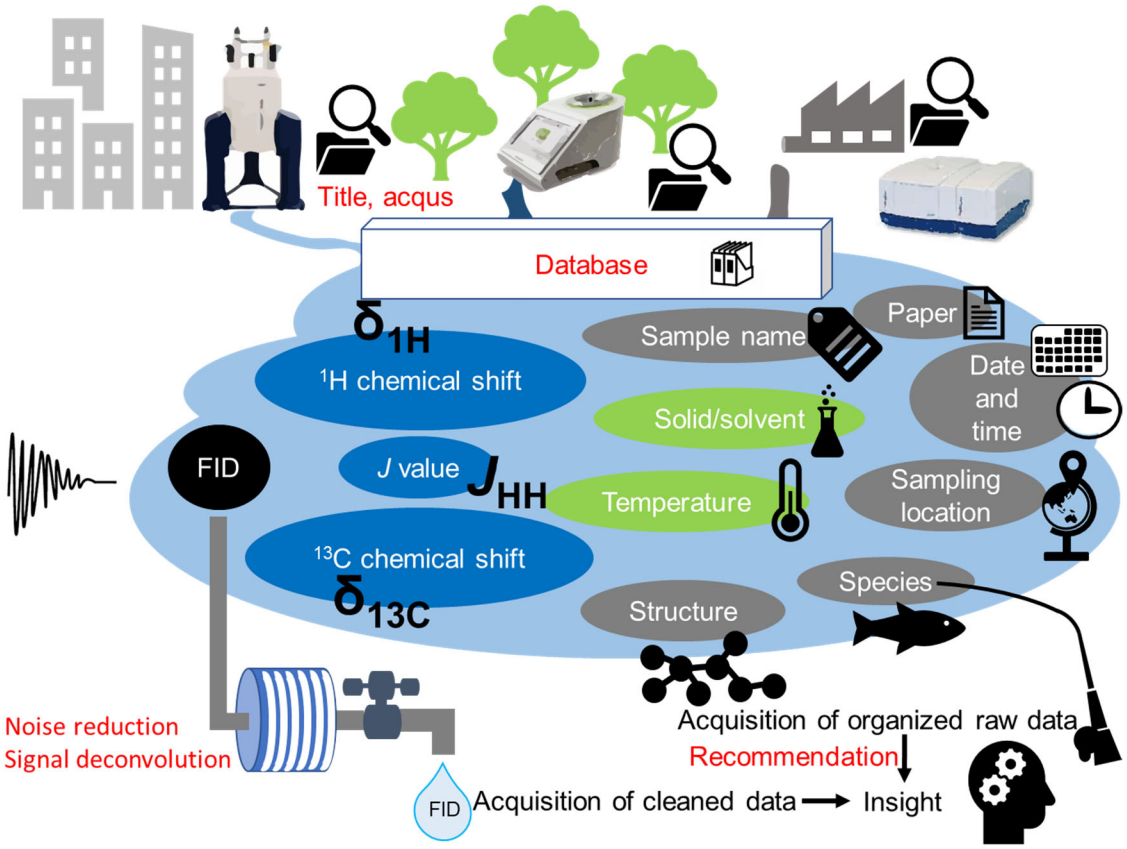


Figure 2. Concept diagram of future plan based on measurement informatics in NMR toward molecular complexity. This figure shows the concept of data lake toward NMR data science. Data lake means the database of NMR data from high-field, low-field, and time-domain NMR. For utilizing NMR data, data cleansing such as noise reduction and signal deconvolution is important toward data-driven analysis.

References

1. Whitesides, G.M.; Ismagilov, R.F. Complexity in chemistry. *Science* **1999**, 284, 89-92.
2. Schmitt-Kopplin, P.; Hemmler, D.; Moritz, F.; Gougeon, R.; Lucio, M.; Meringer, M.; Muller, C.; Harir, M.; Hertkorn, N. Systems chemical analytics: introduction to the challenges of chemical complexity analysis. *Faraday Discussions* **2019**, 218, 9-28.
3. van Beek, T. Low-field benchtop NMR spectroscopy: status and prospects in natural product analysis(dagger). *Phytochemical Analysis* **2021**, 32, 24-37.
4. Takahashi, K.; Tanaka, Y. Materials informatics: a journey towards material design and synthesis. *Dalton Transactions* **2016**, 45, 10497-10499.
5. Kikuchi, J.; Ito, K.; Date, Y. Environmental metabolomics with data science for investigating ecosystem homeostasis. *Prog. Nucl. Magn. Reason. Spectrosc.* **2018**, 104, 56-88.
6. Kikuchi, J.; Yamada, S. NMR window of molecular complexity showing homeostasis in superorganisms. *Analyst* **2017**, 142, 4161-4172.
7. McAlpine, J.; Chen, S.; Kutateladze, A.; MacMillan, J.; Appendino, G.; Barison, A.; Beniddir, M.; Biavatti, M.; Bluml, S.; Boufridi, A., et al. The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. *Natural Product Reports* **2019**, 36, 35-107.
8. Barba, A.; Dominguez, S.; Cobas, C.; Martinsen, D.; Romain, C.; Rzepa, H.; Seoane, F. Workflows Allowing Creation of Journal Article Supporting Information and Findable, Accessible, Interoperable, and Reusable (FAIR)-Enabled Publication of Spectroscopic Data. *Acs Omega* **2019**, 4, 3280-3286.
9. Pupier, M.; Nuzillard, J.M.; Wist, J.; Schlörer, N.E.; Kuhn, S.; Erdelyi, M.; Steinbeck, C.; Williams, A.J.; Butts, C.; Claridge, T.D.W., et al. NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magn. Reason. Chem.* **2018**.
10. Wu, K.; Luo, J.; Zeng, Q.; Dong, X.; Chen, J.; Zhan, C.; Chen, Z.; Lin, Y. Improvement in Signal-to-Noise Ratio of Liquid-State NMR Spectroscopy via a Deep Neural Network DN-Unet. *Anal. Chem.* **2020**.
11. Schlagnitweit, J.; Tang, M.; Baias, M.; Richardson, S.; Schantz, S.; Emsley, L. A solid-state NMR method to determine domain sizes in multi-component polymer formulations. *J. Magn. Res.* **2015**, 261, 43-48.
12. Chen, G.; Shen, Z.; Iyer, A.; Ghumman, U.; Tang, S.; Bi, J.; Chen, W.; Li, Y. Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges. *Polymers* **2020**, 12.
13. Perez-Riverol, Y.; Bai, M.; Leprevost, F.D.; Squizzato, S.; Park, Y.M.; Haug, K.; Carroll, A.J.; Spalding, D.; Paschall, J.; Wang, M.X., et al. Discovering and linking public omics data sets using

the Omics Discovery Index. *Nature Biotechnology* **2017**, 35, 406-409.

14. Wishart, D.S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery* **2016**, 15, 473-484.
15. Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Computational Materials* **2017**, 3, 13.
16. Rosato, A.; Tenori, L.; Cascante, M.; De Atauri Carulla, P.R.; Martins Dos Santos, V.A.P.; Saccenti, E. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics* **2018**, 14, 37.
17. Rodriguez-Martinez, A.; Posma, J.M.; Ayala, R.; Harvey, N.; Jimenez, B.; Neves, A.L.; Lindon, J.C.; Sonomura, K.; Sato, T.A.; Matsuda, F., et al. J-Resolved (1)H NMR 1D-Projections for Large-Scale Metabolic Phenotyping Studies: Application to Blood Plasma Analysis. *Anal. Chem.* **2017**.
18. Wruck, W.; Kashofer, K.; Rehman, S.; Daskalaki, A.; Berg, D.; Gralka, E.; Jozefczuk, J.; Drews, K.; Pandey, V.; Regenbrecht, C., et al. Multi-omic profiles of human non-alcoholic fatty liver disease tissue highlight heterogenic phenotypes. *Sci. Data* **2015**, 2, 150068.
19. Shiokawa, Y.; Date, Y.; Kikuchi, J. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. *Sci. Rep.* **2018**, 8, 8.
20. Sugahara, H.; Odamaki, T.; Fukuda, S.; Kato, T.; Xiao, J.Z.; Abe, F.; Kikuchi, J.; Ohno, H. Probiotic Bifidobacterium longum alters gut luminal metabolism through modification of the gut microbial community. *Sci. Rep.* **2015**, 5, 13548.
21. Kato, T.; Fukuda, S.; Fujiwara, A.; Suda, W.; Hattori, M.; Kikuchi, J.; Ohno, H. Multiple omics uncovers host-gut microbial mutualism during prebiotic fructooligosaccharide supplementation. *DNA Res.* **2014**, 21, 469-480.
22. Mekuchi, M.; Sakata, K.; Yamaguchi, T.; Koiso, M.; Kikuchi, J. Trans-omics approaches used to characterise fish nutritional biorhythms in leopard coral grouper (Plectropomus leopardus). *Sci. Rep.* **2017**, 7, 12.
23. Tomita, S.; Saito, K.; Nakamura, T.; Sekiyama, Y.; Kikuchi, J. Rapid discrimination of strain-dependent fermentation characteristics among Lactobacillus strains by NMR-based metabolomics of fermented vegetable juice. *Plos One* **2017**, 12, 18.
24. Komatsu, T.; Kobayashi, T.; Hatanaka, M.; Kikuchi, J. Profiling Planktonic Biomass Using Element-Specific, Multicomponent Nuclear Magnetic Resonance Spectroscopy. *Environ. Sci. Technol.* **2015**.
25. Wei, F.; Ito, K.; Sakata, K.; Date, Y.; Kikuchi, J. Pretreatment and integrated analysis of spectral data reveal seaweed similarities based on chemical diversity. *Anal. Chem.* **2015**, 87, 2819-2826.
26. Ito, K.; Sakata, K.; Date, Y.; Kikuchi, J. Integrated Analysis of Seaweed Components during Seasonal Fluctuation by Data Mining Across Heterogeneous Chemical Measurements with

Network Visualization. *Anal. Chem.* **2014**, 86, 1098-1105.

27. Ogura, T.; Date, Y.; Masukujane, M.; Coetzee, T.; Akashi, K.; Kikuchi, J. Improvement of physical, chemical, and biological properties of aridisol from Botswana by the incorporation of torrefied biomass. *Sci. Rep.* **2016**, 6, 28011.
28. Watanabe, T.; Shino, A.; Akashi, K.; Kikuchi, J. Chemical profiling of Jatropha tissues under different torrefaction conditions: application to biomass waste recovery. *PLoS One* **2014**, 9, e106893.
29. Mori, T.; Tsuboi, Y.; Ishida, N.; Nishikubo, N.; Demura, T.; Kikuchi, J. Multidimensional High-Resolution Magic Angle Spinning and Solution-State NMR Characterization of C-13-labeled Plant Metabolites and Lignocellulose. *Sci. Rep.* **2015**, 5, 12.
30. Bundy, J.G.; Davey, M.P.; Viant, M.R. Environmental metabolomics: a critical review and future perspectives. *Metabolomics* **2009**, 5, 3-21.
31. Wei, F.F.; Sakata, K.; Asakura, T.; Date, Y.; Kikuchi, J. Systemic Homeostasis in Metabolome, Ionome, and Microbiome of Wild Yellowfin Goby in Estuarine Ecosystem. *Sci. Rep.* **2018**, 8, 12.
32. Ogawa, D.M.O.; Moriya, S.; Tsuboi, Y.; Date, Y.; Prieto-da-Silva, A.R.B.; Radis-Baptista, G.; Yamane, T.; Kikuchi, J. Biogeochemical Typing of Paddy Field by a Data-Driven Approach Revealing Sub-Systems within a Complex Environment - A Pipeline to Filtrate, Organize and Frame Massive Dataset from Multi-Omics Analyses. *Plos One* **2014**, 9, 14.
33. Yoshida, S.; Date, Y.; Akama, M.; Kikuchi, J. Comparative metabolomic and ionomic approach for abundant fishes in estuarine environments of Japan. *Sci. Rep.* **2014**, 4.
34. Asakura, T.; Date, Y.; Kikuchi, J. Comparative Analysis of Chemical and Microbial Profiles in Estuarine Sediments Sampled from Kanto and Tohoku Regions in Japan. *Anal. Chem.* **2014**, 86, 5425-5432.
35. Hashi, K.; Ohki, S.; Matsumoto, S.; Nishijima, G.; Goto, A.; Deguchi, K.; Yamada, K.; Noguchi, T.; Sakai, S.; Takahashi, M., et al. Achievement of 1020 MHz NMR. *J. Magn. Res.* **2015**, 256, 30-33.
36. Halse, M.E. Perspectives for hyperpolarisation in compact NMR. *Trac. Trends Anal. Chem.* **2016**, 83, 76-83.
37. Aslam, N.; Pfender, M.; Neumann, P.; Reuter, R.; Zappe, A.; de Oliveira, F.F.; Denisenko, A.; Sumiya, H.; Onoda, S.; Isoya, J., et al. Nanoscale nuclear magnetic resonance with chemical resolution. *Science* **2017**, 357, 5.
38. Tayler, M.C.D.; Theis, T.; Sjolander, T.F.; Blanchard, J.W.; Kentner, A.; Pustelny, S.; Pines, A.; Budker, D. Invited Review Article: Instrumentation for nuclear magnetic resonance in zero and ultralow magnetic field. *Rev. Sci. Instrum.* **2017**, 88, 17.
39. Blümich, B. Virtual special issue: Magnetic resonance at low fields. *J. Magn. Res.* **2017**, 274, 145-147.

40. Blümich, B.; Singh, K. Desktop NMR and Its Applications From Materials Science To Organic Chemistry. *Angew. Chem. Int. Ed. Engl.* **2018**, 57, 6996-7010.
41. Giraudeau, P.; Massou, S.; Robin, Y.; Cahoreau, E.; Portais, J.C.; Akoka, S. Ultrafast quantitative 2D NMR: an efficient tool for the measurement of specific isotopic enrichments in complex biological mixtures. *Anal. Chem.* **2011**, 83, 3112-3119.
42. Wu, H.; Southam, A.D.; Hines, A.; Viant, M.R. High-throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Anal. Biochem.* **2008**, 372, 204-212.
43. Lin, C.Y.; Wu, H.F.; Tjeerdema, R.S.; Viant, M.R. Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics* **2007**, 3, 55-67.
44. Sekiyama, Y.; Chikayama, E.; Kikuchi, J. Evaluation of a semipolar solvent system as a step toward heteronuclear multidimensional NMR-based metabolomics for 13C-labeled bacteria, plants, and animals. *Anal. Chem.* **2011**, 83, 719-726.
45. Sekiyama, Y.; Chikayama, E.; Kikuchi, J. Profiling polar and semipolar plant metabolites throughout extraction processes using a combined solution-state and high-resolution magic angle spinning NMR approach. *Anal. Chem.* **2010**, 82, 1643-1652.
46. Kim, H.; Ralph, J. Solution-state 2D NMR of ball-milled plant cell wall gels in DMSO-d(6)/pyridine-d(5). *Org. Biomol. Chem.* **2010**, 8, 576-591.
47. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N., et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic. Acids Res.* **2018**, 46, D608-D617.
48. Ulrich, E.L.; Akutsu, H.; Doreleijers, J.F.; Harano, Y.; Ioannidis, Y.E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z., et al. BioMagResBank. *Nucleic. Acids Res.* **2008**, 36, D402-408.
49. Ludwig, C.; Easton, J.M.; Lodi, A.; Tiziani, S.; Manzoor, S.E.; Southam, A.D.; Byrne, J.J.; Bishop, L.M.; He, S.; Arvanitis, T.N., et al. Birmingham Metabolite Library: a publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **2012**, 8, 8-18.
50. Cui, Q.; Lewis, I.A.; Hegeman, A.D.; Anderson, M.E.; Li, J.; Schulte, C.F.; Westler, W.M.; Eghbalnia, H.R.; Sussman, M.R.; Markley, J.L. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **2008**, 26, 162-164.
51. Kuhn, S.; Schlorer, N. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2-a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn. Res. Chem.* **2015**, 53, 582-589.
52. Bingol, K.; Zhang, F.; Bruschweiler-Li, L.; Brüschweiler, R. TOCCATA: a customized carbon total correlation spectroscopy NMR metabolomics database. *Anal. Chem.* **2012**, 84, 9395-9401.
53. Bingol, K.; Li, D.; Zhang, B.; Bruschweiler, 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, 12411-12418.

54. Kale, N.S.; Haug, K.; Conesa, P.; Jayseelan, K.; Moreno, P.; Rocca-Serra, P.; Nainala, V.C.; Spicer, R.A.; Williams, M.; Li, X., et al. MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr. Protoc. Bioinformatics* **2016**, 53, 14.13.11-18.
55. Xia, J.; Wishart, D.S. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. *Curr. Protoc. Bioinformatics* **2016**, 55, 14.10.11-14.10.91.
56. Chikayama, E.; Sekiyama, Y.; Okamoto, M.; Nakanishi, Y.; Tsuboi, Y.; Akiyama, K.; Saito, K.; Shinozaki, K.; Kikuchi, J. Statistical indices for simultaneous large-scale metabolite detections for a single NMR spectrum. *Anal. Chem.* **2010**, 82, 1653-1658.
57. Kikuchi, J.; Tsuboi, Y.; Komatsu, K.; Gomi, M.; Chikayama, E.; Date, Y. SpinCouple: Development of a Web Tool for Analyzing Metabolite Mixtures via Two-Dimensional J-Resolved NMR Database. *Anal. Chem.* **2016**, 88, 659-665.
58. Kikuchi, J.; Ogata, Y.; Shinozaki, K. ECOMICS: ECosytem trans-OMICS tools and methods for complex environmental samples and datasets. *J. Ecosys. Ecogr.* **2011**, S2, 001.
59. Dashti, H.; Westler, W.M.; Tonelli, M.; Wedell, J.R.; Markley, J.L.; Eghbalnia, H.R. Spin System Modeling of Nuclear Magnetic Resonance Spectra for Applications in Metabolomics and Small Molecule Screening. *Anal. Chem.* **2017**, 89, 12201-12208.
60. Dashti, H.; Wedell, J.R.; Westler, W.M.; Tonelli, M.; Aceti, D.; Amarasinghe, G.K.; Markley, J.L.; Eghbalnia, H.R. Applications of Parametrized NMR Spin Systems of Small Molecules. Anal Chem 2018, 90, 10646-10649.
61. Toumi, I.; Caldarelli, S.; Torrésani, B. A review of blind source separation in NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, 81, 37-64.
62. Toumi, I.; Torrésani, B.; Caldarelli, S. Effective processing of pulse field gradient NMR of mixtures by blind source separation. *Anal. Chem.* **2013**, 85, 11344-11351.
63. Misawa, T.; Komatsu, T.; Date, Y.; Kikuchi, J. SENSI: signal enhancement by spectral integration for the analysis of metabolic mixtures. *Chem. Commun.* **2016**, 52, 2964-2967.
64. Bouveresse, D.J.R.; Moya-Gonzalez, A.; Ammari, F.; Rutledge, D.N. Two novel methods for the determination of the number of components in independent components analysis models. *Chemom. Intell. Lab. Syst.* **2012**, 112, 24-32.
65. Chikayama, E.; Yamashina, R.; Komatsu, K.; Tsuboi, Y.; Sakata, K.; Kikuchi, J.; Sekiyama, Y. FoodPro: A Web-Based Tool for Evaluating Covariance and Correlation NMR Spectra Associated with Food Processes. *Metabolites* **2016**, 6.
66. Komatsu, T.; Kobayashi, T.; Hatanaka, M.; Kikuchi, J. Profiling Planktonic Biomass Using Element-Specific, Multicomponent Nuclear Magnetic Resonance Spectroscopy. *Env. Sci. Technol.* **2015**, 49, 7056-7062.
67. Takeuchi, K.; Baskaran, K.; Arthanari, H. Structure determination using solution NMR: Is it worth

the effort? *J. Magn. Reson.* **2019**, 306, 195-201.

68. Jimenez, B.; Holmes, E.; Heude, C.; Tolson, R.; Harvey, N.; Lodge, S.; Chetwynd, A.; Cannet, C.; Fang, F.; Pearce, J., et al. Quantitative Lipoprotein Subclass and Low Molecular Weight Metabolite Analysis in Human Serum and Plasma by H-1 NMR Spectroscopy in a Multilaboratory Trial. *Anal. Chem.* **2018**, 90, 11962-11971.
69. Singh, K.; Kumar, S.P.; Blumich, B. Monitoring the mechanism and kinetics of a transesterification reaction for the biodiesel production with low field H-1 NMR spectroscopy. *Fuel* **2019**, 243, 192-201.
70. Wishart, D.S. NMR metabolomics: A look ahead. *J. Magn. Reason.* **2019**, 306, 155-161.
71. Maeda, H.; Yanagisawa, Y. Future prospects for NMR magnets: A perspective. *J. Magn. Reason.* **2019**, 306, 80-85.
72. Kovacs, H.; Moskau, D.; Spraul, M. Cryogenically cooled probes - a leap in NMR technology. *Prog. Nucl. Magn. Reason. Spectrosc.* **2005**, 46, 131-155.
73. Clos, L.J.; Jofre, M.F.; Ellinger, J.J.; Westler, W.M.; Markley, J.L. NMRbot: Python scripts enable high-throughput data collection on current Bruker BioSpin NMR spectrometers. *Metabolomics* **2013**, 9, 558-563.
74. Ardenkjaer-Larsen, J.H.; Fridlund, B.; Gram, A.; Hansson, G.; Hansson, L.; Lerche, M.H.; Servin, R.; Thaning, M.; Golman, K. Increase in signal-to-noise ratio of > 10,000 times in liquid-state NMR. *Proc. Natl. Acad. Sci. USA* **2003**, 100, 10158-10163.
75. Kazimierczuk, K.; Orekhov, V. Non-uniform sampling: post-Fourier era of NMR data collection and processing. *Magn. Reason. Chem.* **2015**, 53, 921-926.
76. Pines, A.; Waugh, J.S.; Gibby, M.G. Proton-enhanced nuclear induction spectroscopy. A method for high resolution NMR of dilute spins in solids. *J. Chem. Phys.* **1972**, 56, 1776-1777.
77. Morris, G.A.; Freeman, R. Enhancement of nuclear magnetic resonance signals by polarization transfer. *J. Am. Chem. Soc.* **1979**, 101, 760-762.
78. Blumich, B. Low-field and benchtop NMR. *J. Magn. Reason.* **2019**, 306, 27-35.
79. Meiboom, S.; Gill, D. Modified spin-echo method for measuring nuclear relaxation times. *Rev. Sci. Instrum.* **1958**, 29, 688-691.
80. Piotto, M.; Saudek, V.; Sklenár, V. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **1992**, 2, 661-665.
81. Vilen, E.M.; Klinger, M.; Sandstrom, C. Application of diffusion-edited NMR spectroscopy for selective suppression of water signal in the determination of monomer composition in alginates. *Magn. Reason. Chem.* **2011**, 49, 584-591.
82. Chandrakumar, N. 1D Double Quantum Filter NMR Studies. *Annu. Rep. NMR Spectrosc.* **2009**, 67, 265-329.
83. Lopez, J.M.; Cabrera, R.; Maruenda, H. Ultra-Clean Pure Shift 1H-NMR applied to metabolomics

profiling. *Sci. Rep.* **2019**, 9.

84. Gouilleux, B.; Rouger, L.; Giraudeau, P. Ultrafast 2D NMR: Methods and Applications. *Annu. Rep. NMR Spectrosc.* **2018**, 93, 75-144.
85. Castanar, L.; Dal Poggetto, G.; Colbourne, A.; Morris, G.; Nilsson, M. The GNAT: A new tool for processing NMR data. *Magn. Reason. Chem.* **2018**, 56, 546-558.
86. Morris, G.; Barjat, H.; Horne, T. Reference deconvolution methods. *Prog. Nucl. Magn. Reason. Spectrosc.* **1997**, 31, 197-257.
87. Taylor, H.; Haiges, R.; Kershaw, A. Increasing Sensitivity in Determining Chemical Shifts in One Dimensional Lorentzian NMR Spectra. *J. Phys. Chem. A* **2013**, 117, 3319-3331.
88. Krishnamurthy, K. CRAFT (complete reduction to amplitude frequency table) - robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR. *Magn. Reason. Chem.* **2013**, 51, 821-829.
89. Ibrahim, M.; Pardi, C.; Brown, T.; McDonald, P. Active elimination of radio frequency interference for improved signal-to-noise ratio for in-situ NMR experiments in strong magnetic field gradients. *J. Magn. Reason.* **2018**, 287, 99-109.
90. Langmead, C.J.; Donald, B.R. Extracting structural information using time-frequency analysis of protein NMR data. In Proceedings of RECOMB 2001: proceedings of the Fifth Annual International Conference on Computational Biology; pp. 164–175.
91. Hirakawa, K.; Koike, K.; Kanawaku, Y.; Moriyama, T.; Sato, N.; Suzuki, T.; Furihata, K.; Ohno, Y. Short-time Fourier Transform of Free Induction Decays for the Analysis of Serum Using Proton Nuclear Magnetic Resonance. *J. Oleo Sci.* **2019**, 68, 369-378.
92. Short, T.; Alzapiedi, L.; Bruschweiler, R.; Snyder, D. A Covariance NMR Toolbox for MATLAB and OCTAVE. *J. Magn. Reason.* **2011**, 209, 75-78.
93. Manu, V.; Gopinath, T.; Wang, S.; Veglia, G. T-2* weighted Deconvolution of NMR Spectra: Application to 2D Homonuclear MAS Solid-State NMR of Membrane Proteins. *Sci. Rep.* **2019**, 9.
94. Yamada, S.; Ito, K.; Kurotani, A.; Yamada, Y.; Chikayama, E.; Kikuchi, J. InterSpin: Integrated Supportive Webtools for Low- and High-Field NMR Analyses Toward Molecular Complexity. *Acs Omega* **2019**, 4, 3361-3369.
95. Kusaka, Y.; Hasegawa, T.; Kaji, H. Noise Reduction in Solid-State NMR Spectra Using Principal Component Analysis. *J. Phys. Chem. A* **2019**, 123, 10333-10338.
96. Stilbs, P. Automated CORE, RECORD, and GRECORD processing of multi-component PGSE NMR diffusometry data. *Eur. Biophys. J. Biophys. Lett.* **2013**, 42, 25-32.
97. Stilbs, P. RECORD processing - A robust pathway to component-resolved HR-PGSE NMR diffusometry. *J. Magn. Reason.* **2010**, 207, 332-336.
98. Stilbs, P.; Paulsen, K.; Griffiths, P. Global least-squares analysis of large, correlated spectral data sets: Application to component-resolved FT-PGSE NMR spectroscopy. *J. Phys. Chem.* **1996**, 100,

8180-8189.

99. Halouska, S.; Powers, R. Negative impact of noise on the principal component analysis of NMR data. *J. Magn. Reason.* **2006**, 178, 88-95.

100. Becker, E.; Ferretti, J.; Gambhir, P. Selection of optimum parameters for pulse Fourier transform nuclear magnetic resonance. *Anal. Chem.* **1979**, 51, 1413-1420.

101. Mo, H.; Harwood, J.; Zhang, S.; Xue, Y.; Santini, R.; Raftery, D. R: A quantitative measure of NMR signal receiving efficiency. *J. Magn. Reason.* **2009**, 200, 239-244.

102. Mo, H.P.; Harwood, J.S.; Raftery, D. A quick diagnostic test for NMR receiver gain compression. *Magn. Reason. Chem.* **2010**, 48, 782-786.

103. Mo, H.; Harwood, J.S.; Raftery, D. Receiver gain function: the actual NMR receiver gain. *Magn. Reason. Chem.* **2010**, 48, 235-238.

104. Mo, H.P.; Harwood, J.; Raftery, D. NMR quantitation: influence of RF inhomogeneity. *Magn. Reason. Chem.* **2011**, 49, 655-658.

105. Liu, H.; Dong, H.; Ge, J.; Bai, B.; Yuan, Z.; Zhao, Z. Research on a secondary tuning algorithm based on SVD & STFT for FID signal. *Meas. Sci. Technol.* **2016**, 27.

106. Zitnik, M.; Zupan, B. NIMFA: A Python Library for Nonnegative Matrix Factorization. *J. Mach. Learn. Res.* **2012**, 13, 849-853.

107. Liu, H.; Dong, H.; Ge, J.; Liu, Z.; Yuan, Z.; Zhu, J.; Zhang, H. A fusion of principal component analysis and singular value decomposition based multivariate denoising algorithm for free induction decay transversal data. *Rev. Sci. Instrum.* **2019**, 90.

108. Keeler, J. Understanding NMR Spectroscopy (2004). Availabe online: https://doi.org/10.17863/CAM.1291 (accessed on 10 February 2021).

109. Dueck, D.; Morris, Q.; Frey, B. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* **2005**, 21, I144-I151.

110. Claridge, T. MNova: NMR Data Processing, Analysis, and Prediction Software. *J. Chem. Inf. Model.* **2009**, 49, 1136-1137.

111. Larive, C.K., Jayawickrama, D., Orfi, L. Quantitative analysis of peptides with NMR spectroscopy. *Appl. Spectrosc.* **1997**, 51(10), 1531-1536.

112. Helmus, J.; Jaroniec, C. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *J. Biomol. NMR* **2013**, 55, 355-367.

113. Laurberg, H.; Christensen, M.G.; Plumbley, M.D.; Hansen, L.K.; Jensen, S.H. Theorems on positive data: on the uniqueness of NMF. *Comput. Intell. Neurosci.* **2008**, 764206.

114. Kim, H.; Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **2007**, 23, 1495-1502.

115. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, 401, 788-791.

116. Demchak, B.; Hull, T.; Reich, M.; Liefeld, T.; Smoot, M.; Ideker, T.; Mesirov, J.P. Cytoscape: the network visualization tool for GenomeSpace workflows. *F1000Res* **2014**, 3, 151.

117. Misawa, T.; Wei, F.; Kikuchi, J. Application of Two-Dimensional Nuclear Magnetic Resonance for Signal Enhancement by Spectral Integration Using a Large Data Set of Metabolic Mixtures. *Anal. Chem.* **2016**, 88, 6130-6134.

118. Asakura, T.; Sakata, K.; Date, Y.; Kikuchi, J. Regional feature extraction of various fishes based on chemical and microbial variable selection using machine learning. *Anal. Methods* **2018**, 10, 2160-2168.

119. Wei, F.; Fukuchi, M.; Ito, K.; Sakata, K.; Asakura, T.; Date, Y.; Kikuchi, J. Large-Scale Evaluation of Major Soluble Macromolecular Components of Fish Muscle from a Conventional H-1-NMR Spectral Database. *Molecules* **2020**, 25.

120. Hepburn, C.; Adlen, E.; Beddington, J.; Carter, E.A.; Fuss, S.; Mac Dowell, N.; Minx, J.C.; Smith, P.; Williams, C.K. The technological and economic prospects for CO2 utilization and removal. *Nature* **2019**, 575, 87-97.

121. Zhu, Y.; Romain, C.; Williams, C.K. Sustainable polymers from renewable resources. *Nature* **2016**, 540, 354-362.

122. Inkinen, S.; Hakkarainen, M.; Albertsson, A.; Sodergard, A. From Lactic Acid to Poly(lactic acid) (PLA): Characterization and Analysis of PLA and Its Precursors. *Biomacromolecules* **2011**, 12, 523-532.

123. Schaler, K.; Achilles, A.; Barenwald, R.; Hackel, C.; Saalwachter, K. Dynamics in Crystallites of Poly(epsilon-caprolactone) As Investigated by Solid-State NMR. *Macromolecules* **2013**, 46, 7818-7825.

124. Foston, M. Advances in solid-state NMR of cellulose. *Curr. Opin. Biotechnol.* **2014**, 27, 176-184.

125. Okushita, K.; Chikayama, E.; Kikuchi, J. Solubilization mechanism and characterization of the structural change of bacterial cellulose in regenerated states through ionic liquid treatment. *Biomacromolecules* **2012**, 13, 1323-1330.

126. Mori, T.; Chikayama, E.; Tsuboi, Y.; Ishida, N.; Shisa, N.; Noritake, Y.; Moriya, S.; Kikuchi, J. Exploring the conformational space of amorphous cellulose using NMR chemical shifts. *Carbohydr. Polym.* **2012**, 90, 1197-1203.

127. Komatsu, T.; Kikuchi, J. Selective Signal Detection in Solid-State NMR Using Rotor-Synchronized Dipolar Dephasing for the Analysis of Hemicellulose in Lignocellulosic Biomass. *J. Phys. Chem. Lett.* **2013**, 4, 2279-2283.

128. Okushita, K.; Komatsu, T.; Chikayama, E.; Kikuchi, J. Statistical approach for solid-state NMR spectra of cellulose derived from a series of variable parameters. *Polym. J.* **2012**, 44, 895-900.

129. Yamazawa, A.; Iikura, T.; Shino, A.; Date, Y.; Kikuchi, J. Solid-, Solution-, and Gas-state NMR Monitoring of C-13-Cellulose Degradation in an Anaerobic Microbial Ecosystem. *Molecules* **2013**,

18, 9021-9033.

130. Komatsu, T.; Kikuchi, J. Comprehensive signal assignment of 13C-labeled lignocellulose using multidimensional solution NMR and 13C chemical shift comparison with solid-state NMR. *Anal. Chem.* **2013**, 85, 8857-8865.

131. Yamazawa, A.; Iikura, T.; Morioka, Y.; Shino, A.; Ogata, Y.; Date, Y.; Kikuchi, J. Cellulose digestion and metabolism induced biocatalytic transitions in anaerobic microbial ecosystems. *Metabolites* **2013**, 4, 36-52.

132. Eden, M. Editorial for the Special Issue on Solid-State NMR Spectroscopy in Materials Chemistry. *Molecules* **2020**, 25.

133. Wang, Z.; Hanrahan, M.; Kobayashi, T.; Perras, F.; Chen, Y.; Engelke, F.; Reiter, C.; Purea, A.; Rossini, A.; Pruski, M. Combining fast magic angle spinning dynamic nuclear polarization with indirect detection to further enhance the sensitivity of solid-state NMR spectroscopy. *Solid State Nucl. Magn. Reason.* **2020**, 109.

134. Pustovalova, Y.; Hoch, J. Sensetivity Gain in Nonuniformly Sampled NMR Experiments. *Biophys. J.* **2020**, 118, 612A-612A.

135. Sugishita, T.; Matsuki, Y.; Fujiwara, T. Absolute H-1 polarization measurement with a spin-correlated component of magnetization by hyperpolarized MAS-DNP solid-state NMR. *Solid State Nucl. Magn. Reason.* **2019**, 99, 20-26.

136. Plainchont, B.; Berruyer, P.; Dumez, J.; Tannin, S.; Giraudeau, P. Dynamic Nuclear Polarization Opens New Perspectives for NMR Spectroscopy in Analytical Chemistry. *Anal. Chem.* **2018**, 90, 3639-3650.

137. Chen, K. A Practical Review of NMR Lineshapes for Spin-1/2 and Quadrupolar Nuclei in Disordered Materials. *Int. J. Mol. Sci.* **2020**, 21.

138. Demco, D.; Johansson, A.; Tegenfeldt, J. Proton spin diffusion for spatial heterogeneity and morphology investigations of polymers. *Solid State Nucl. Magn. Reason.* **1995**, 4, 13-38.

139. Buda, A.; Demco, D.; Bertmer, M.; Blumich, B.; Reining, B.; Keul, H.; Hocker, H. Domain sizes in heterogeneous polymers by spin diffusion using single-quantum and double-quantum dipolar filters. *Solid State Nucl. Magn. Reason.* **2003**, 24, 39-67.

140. Masuda, K.; Kaji, H.; Horii, F. Solid-state C-13 NMR and H-1 CRAMPS investigations of the hydration process and hydrogen bonding for poly(vinyl alcohol) films. *Polym. J.* **2001**, 33, 356-363.

141. Struppe, J.; Quinn, C.; Lu, M.; Wang, M.; Hou, G.; Lu, X.; Kraus, J.; Andreas, L.; Stanek, J.; Lalli, D., et al. Expanding the horizons for structural analysis of fully protonated protein assemblies by NMR spectroscopy at MAS frequencies above 100 kHz. *Solid State Nucl. Magn. Reason.* **2017**, 87, 117-125.

142. Besghini, D.; Mauri, M.; Simonutti, R. Time Domain NMR in Polymer Science: From the

Laboratory to the Industry. *Appl. Sci.* **2019**, 9.

143. Ogura, T.; Date, Y.; Kikuchi, J. Differences in Cellulosic Supramolecular Structure of Compositionally Similar Rice Straw Affect Biomass Metabolism by Paddy Soil Microbiota. *Plos One* **2013**, 8.
144. Mileo, P.; Yuan, S.; Ayala, S.; Duan, P.; Semino, R.; Cohen, S.; Schmidt-Rohr, K.; Maurin, G. Structure of the Polymer Backbones in polyMOF Materials. *J. Am. Chem. Soc.* **2020**, 142, 10863-10868.
145. Schaler, K.; Roos, M.; Micke, P.; Golitsyn, Y.; Seidlitz, A.; Thurn-Albrecht, T.; Schneider, H.; Hempel, G.; Saalwachter, K. Basic principles of static proton low-resolution spin diffusion NMR in nanophase-separated materials with mobility contrast. *Solid State Nucl. Magn. Reason.* **2015**, 72, 50-63.
146. Schneider, H.; Saalwachter, K.; Roos, M. Complex Morphology of the Intermediate Phase in Block Copolymers and Semicrystalline Polymers As Revealed by H-1 NMR Spin Diffusion Experiments. *Macromolecules* **2017**, 50, 8598-8610.
147. Weingarth, M.; Tekely, P.; Bruschweiler, R.; Bodenhausen, G. Improving the quality of 2D solid-state NMR spectra of microcrystalline proteins by covariance analysis. *Chem. Commun.* **2010**, 46, 952-954.
148. Bak, M.; Rasmussen, J.; Nielsen, N. SIMPSON: A general simulation program for solid-state NMR spectroscopy. *J. Magn. Reason.* **2000**, 147, 296-330.
149. Veshtort, M.; Griffin, R. SPINEVOLUTION: A powerful tool for the simulation of solid and liquid state NMR experiments. *J. Magn. Reason.* **2006**, 178, 248-282.
150. Massiot, D.; Fayon, F.; Capron, M.; King, I.; Le Calve, S.; Alonso, B.; Durand, J.; Bujoli, B.; Gan, Z.; Hoatson, G. Modelling one- and two-dimensional solid-state NMR spectra. *Magn. Reason. Chem.* **2002**, 40, 70-76.
151. Grimminck, D.; van Meerten, B.; Verkuijlen, M.; van Eck, E.; Meerts, W.; Kentgens, A. EASY-GOING deconvolution: Automated MQMAS NMR spectrum on a model with analytical crystallite excitation efficiencies. *J. Magn. Reason.* **2013**, 228, 116-124.
152. Smith, A. INFOS: spectrum fitting software for NMR analysis. *J. Biomol. NMR* **2017**, 67, 77-94.
153. Wojdyr, M. Fityk: a general-purpose peak fitting program. *J. Appl. Cryst.* **2010**, 43, 1126-1128.
154. van Meerten, S.; Franssen, W.; Kentgens, A. ssNake: A cross-platform open-source NMR data processing and fitting application. *J. Magn. Reason.* **2019**, 301, 56-66.
155. Yamada, S.; Kurotani, A.; Chikayama, E.; Kikuchi, J. Signal Deconvolution and Noise Factor Analysis Based on a Combination of Time-Frequency Analysis and Probabilistic Sparse Matrix Factorization. *Int. J. Mol. Sci.* **2020**, 21.
156. Kaneko, H. Data Visualization, Regression, Applicability Domains and Inverse Analysis Based on Generative Topographic Mapping. *Mol. Inf.* **2019**, 38.

157. Kossaifi, J.; Panagakis, Y.; Anandkumar, A.; Pantic, M. TensorLy: Tensor Learning in Python. *J. Mach. Learn. Res.* **2019**, 20.

158. Dal Poggetto, G.; Castanar, L.; Adams, R.; Morris, G.; Nilsson, M. Dissect and Divide: Putting NMR Spectra of Mixtures under the Knife. *J. Am. Chem. Soc.* **2019**, 141, 5766-5771.

159. Kasai, T.; Ono, S.; Koshiba, S.; Yamamoto, M.; Tanaka, T.; Ikeda, S.; Kigawa, T. Amino-acid selective isotope labeling enables simultaneous overlapping signal decomposition and information extraction from NMR spectra. *J. Biomol. NMR* **2020**, 74, 125-137.

160. Ito, K.; Obuchi, Y.; Chikayama, E.; Date, Y.; Kikuchi, J. Exploratory machine-learned theoretical chemical shifts can closely predict metabolic mixture signals. *Chem. Sci.* **2018**, 9, 8213-8220.

161. Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). *J. Chem. Inf. Model.* **2016**, 56, 286-299.

162. Aursand, M.; Standal, I.; Axelson, D. High-resolution (13)C nuclear magnetic resonance spectroscopy pattern recognition of fish oil capsules. *J. Agric. Food Chem.* **2007**, 55, 38-47.

163. Zhang, J.; Terayama, K.; Sumita, M.; Yoshizoe, K.; Ito, K.; Kikuchi, J.; Tsuda, K. NMR-TS: de novo molecule identification from NMR spectra. *Sci. Technol. Adv. Mater.* **2020**, 21, 552-561.

164. Aguilera-Saez, L.; Arrabal-Campos, F.; Callejon-Ferre, A.; Medina, M.; Fernandez, I. Use of multivariate NMR analysis in the content prediction of hemicellulose, cellulose and lignin in greenhouse crop residues. *Phytochemistry* **2019**, 158, 110-119.

165. Tang, Q.; Chen, Y.; Yang, H.; Liu, M.; Xiao, H.; Wu, Z.; Chen, H.; Naqvi, S. Prediction of Bio-oil Yield and Hydrogen Contents Based on Machine Learning Method: Effect of Biomass Compositions and Pyrolysis Conditions. *Energy Fuels* **2020**, 34, 11050-11060.

166. Kasmuri, N.; Kamarudin, S.; Abdullah, S.; Hasan, H.; Som, A. Integrated advanced nonlinear neural network-simulink control system for production of bio-methanol from sugar cane bagasse via pyrolysis. *Energy* **2019**, 168, 261-272.

167. Yucel, O.; Aydin, E.; Sadikoglu, H. Comparison of the different artificial neural networks in prediction of biomass gasification products. *Int. J. Energy Res.* **2019**, 43, 5992-6003.

168. Chen, X.; Zhang, H.; Song, Y.; Xiao, R. Prediction of product distribution and bio-oil heating value of biomass fast pyrolysis. *Chem. Eng. Process. Process Intensif.* **2018**, 130, 36-42.

169. Verma, R.P.; Hansch, C. Use of 13C NMR chemical shift as QSAR/QSPR descriptor. *Chem. Rev.* **2011**, 111, 2865-2899.

170. Himanen, L.; Geurts, A.; Foster, A.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives (vol 6, 1900808, 2019). *Adv. Sci.* **2020**, 7.

171. Ma, R.; Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *J. Chem. Inf. Model.* **2020**, 60, 4684-4690.

172. Granda, J.M.; Donina, L.; Dragone, V.; Long, D.L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, 559, 377-381.

173. Li, M.; Pu, Y.; Chen, F.; Ragauskas, A.J. Synthesis and Characterization of Lignin-grafted-poly(ε-caprolactone) from Different Biomass Sources. *N. Biotechnol.* **2021**, 60, 189-199.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor, Head of Environmental Metabolic Analysis Research team of RIKEN, Visiting Professor at Nagoya University, and Visiting Professor at Yokohama City University Dr. Jun Kikuchi, for his excellent guidance and continuous supports throughout my doctoral research. He strongly inspired me to challenge measurement informatics research, which provides a bird's eye view of molecular complexity through the NMR window, and helped me to establish philosophy and originality for my research. Furthermore, I am also thankful to Professor at the Niigata University of International and Information Studies Dr. Eisuke Chikayama who made active discussions and comments about my measurement informatics approaches in NMR. And, I am deeply thanks to Visiting Researcher at Yokohama City University Dr. Kengo Ito, and Researcher at RIKEN Dr. Atsushi Kurotani for their helpful advices and technical supports during my study period. I wish to thank all members of RIKEN, for sample preparation, NMR measurements.

Secondly, I am very grateful to Visiting Professor Jun Kikuchi for offering me a part-time position in 2017 and a part-time position in the Junior Research Associate (JRA) program from 2018-2020. And, I am very thankful to Nagoya University for the financial supports for tuition fee (half exemption) from 2018-2019. Moreover, I gratefully acknowledge the Nuclear Magnetic Resonance Society of Japan (NMRJ) and the 8th Asia-Pacific NMR Symposium (APNMR) Singapore Committee for travel grants to the European Congress on Magnetic Resonance (EUROMAR) 2017 (Poland on 2-6 July 2017) and APNMR 2019 (Singapore on 3-6 July 2019). Furthermore, All works were partially supported by a grant from the Agriculture, Forestry, and Fisheries Research Council, as well as the Strategic Innovation Program (SIP) from Japanese Cabinet Office (CAO) to Visiting Professor Dr. Jun Kikuchi.

Finally, I would like to thank my family members and friends for their never-ending and unconditional supports even in the age of a global pandemic. I will contribute to my family as well as to future society throughout my life.

March 2021

Shunji Yamada

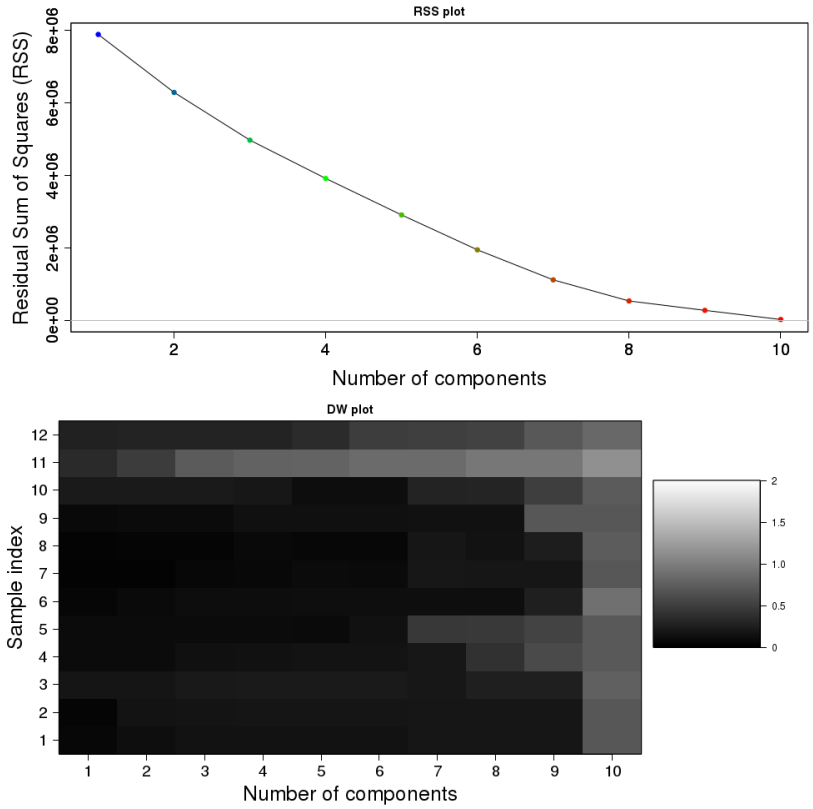
Appendix A

Supporting Information for

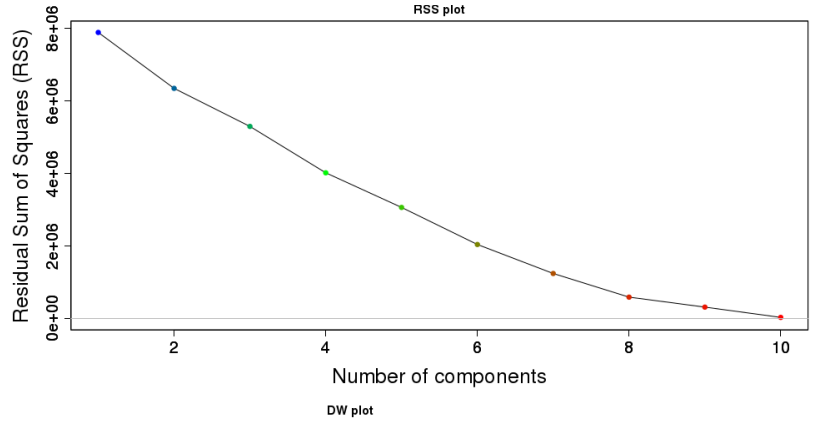
InterSpin: Integrated Supportive Webtools for Low- and High-Field NMR Analyses Toward Molecular Complexity

InterSpin is available at http://dmar.riken.jp/interspin/

a) Fast ICA



b) MCR-ALS



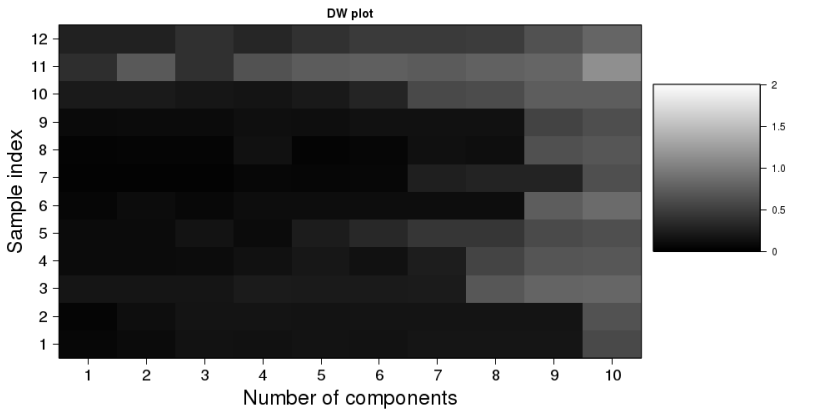
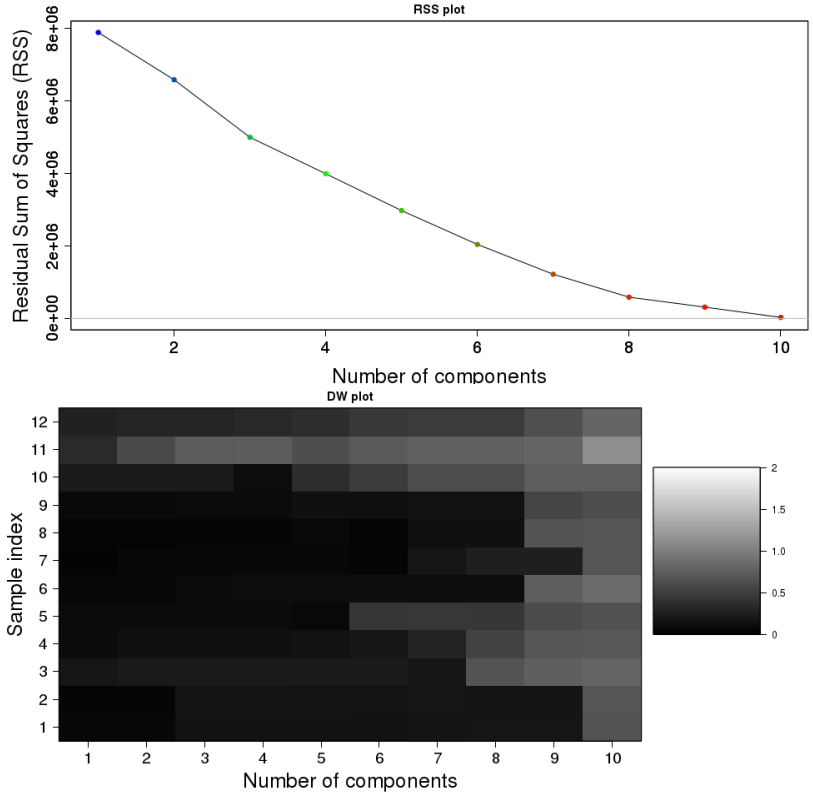


Figure S1. Determining the number of components by RSS and DW plots. (**a**) Fast ICA, and (**b**) MCR-ALS.

c) NMF



d) NNSC

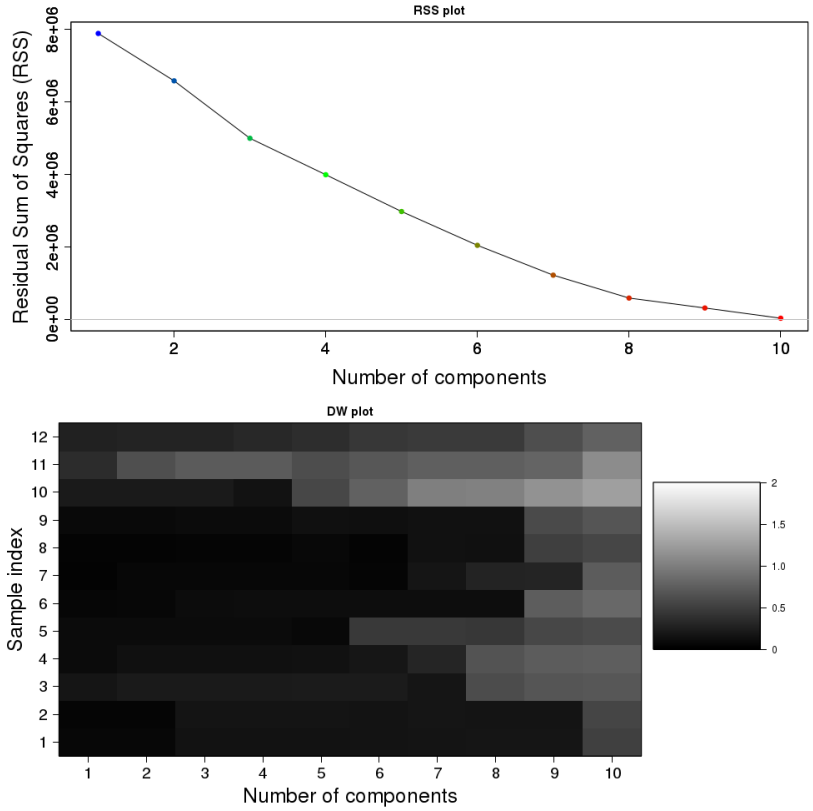


Figure S1. Determining the number of components by RSS and DW plots (continuation). (**c**) NMF, and (**d**) NNSC.

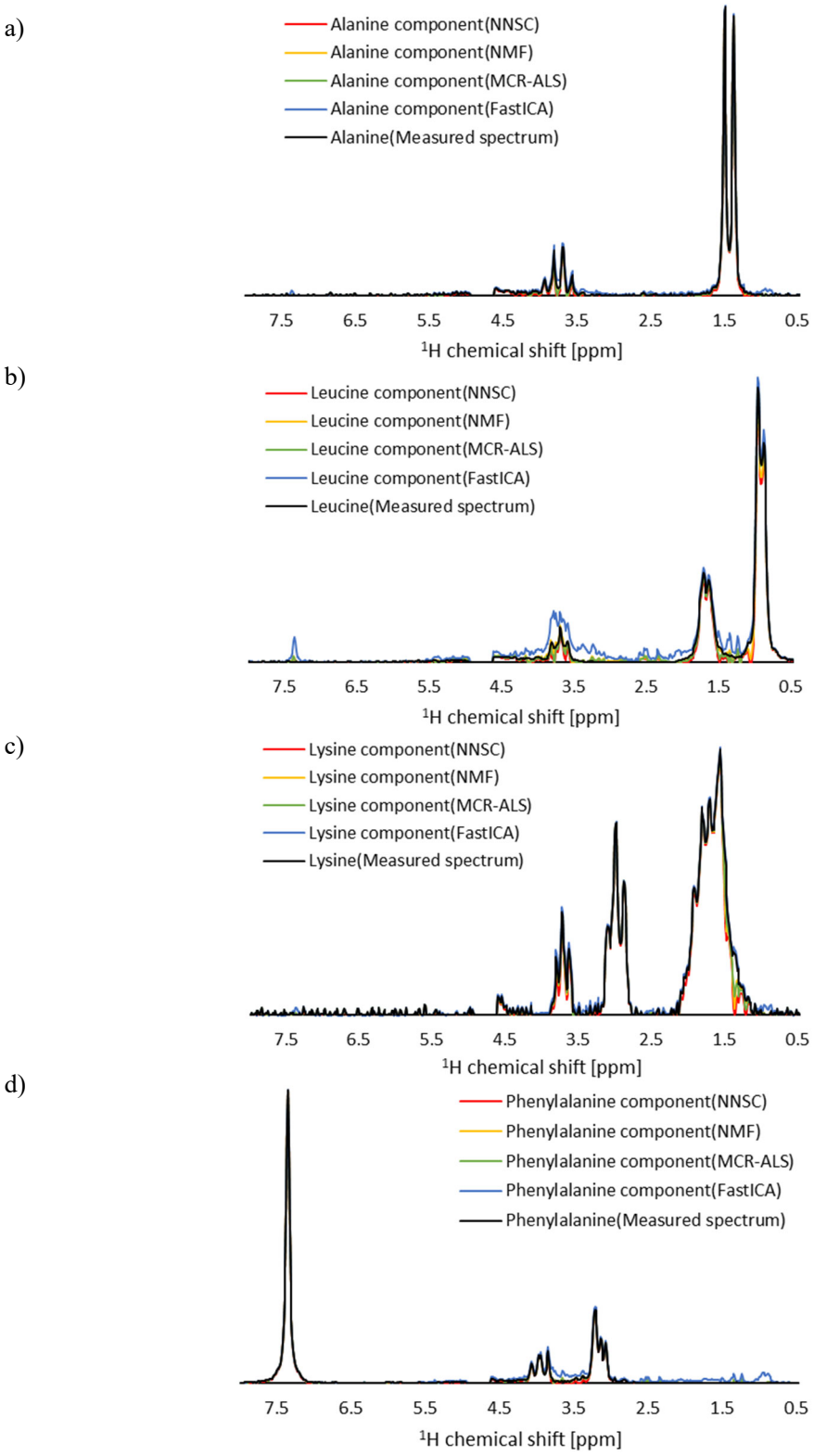
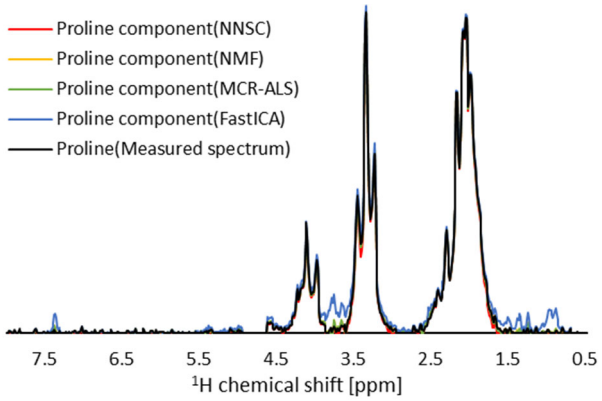
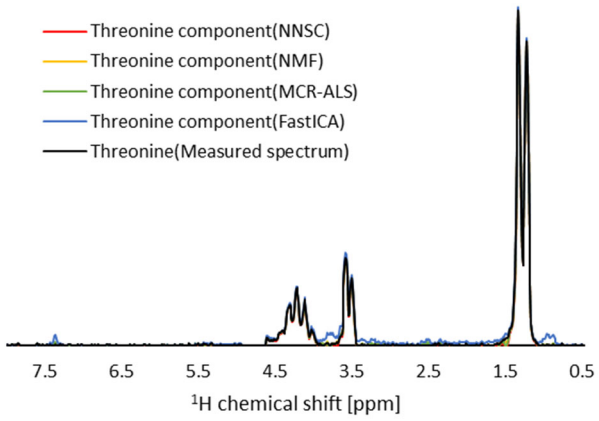


Figure S2. Comparison of the component spectrum separated by different algorithms and the standard spectrum. (**a**) Alanine, (**b**) leucine, (**c**) lysine, and (**d**) phenylalanine.

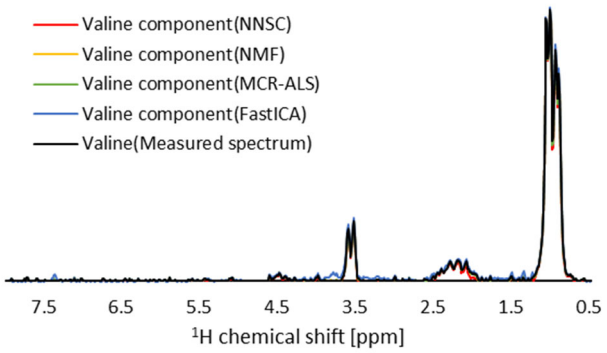
e)



f)



g)



h)

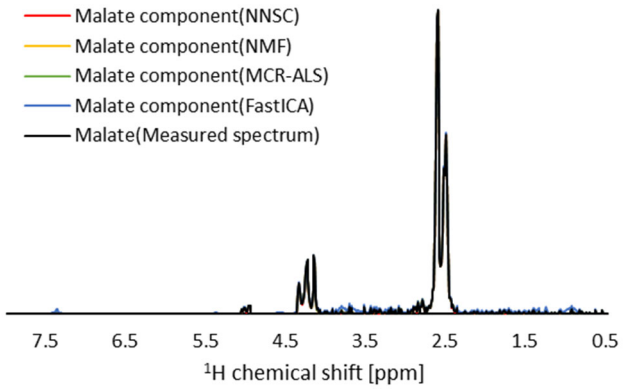
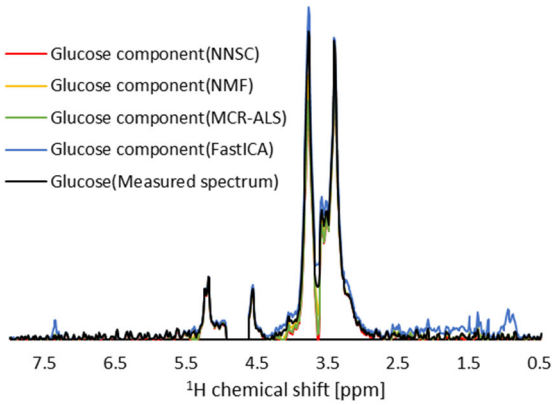


Figure S2. Comparison of the component spectrum separated by different algorithms and the standard spectrum (continuation). (**e**) Proline, (**f**) threonine, (**g**) valine, and (**h**) malate.

i)



j)

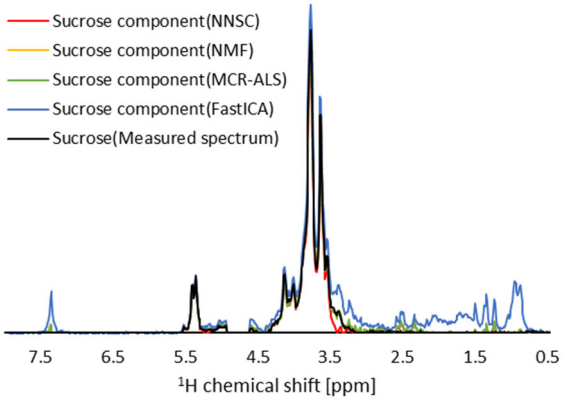


Figure S2. Comparison of the component spectrum separated by different algorithms and the standard spectrum (continuation). (**i**) Glucose, and (**j**) sucrose.

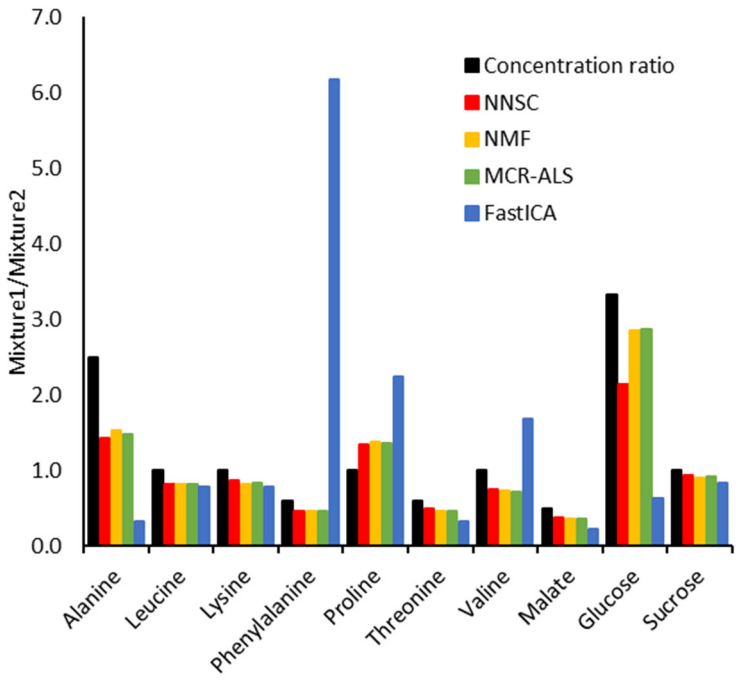


Figure S3. Comparison of concentration ratios and separated spectral score ratios of mixtures 1 and 2. Shown are the concentration ratio of mixtures 1 and 2, and the score ratio of the spectrum separated by each algorithm (NNSC, NMF, MCR-ALS, and Fast ICA).

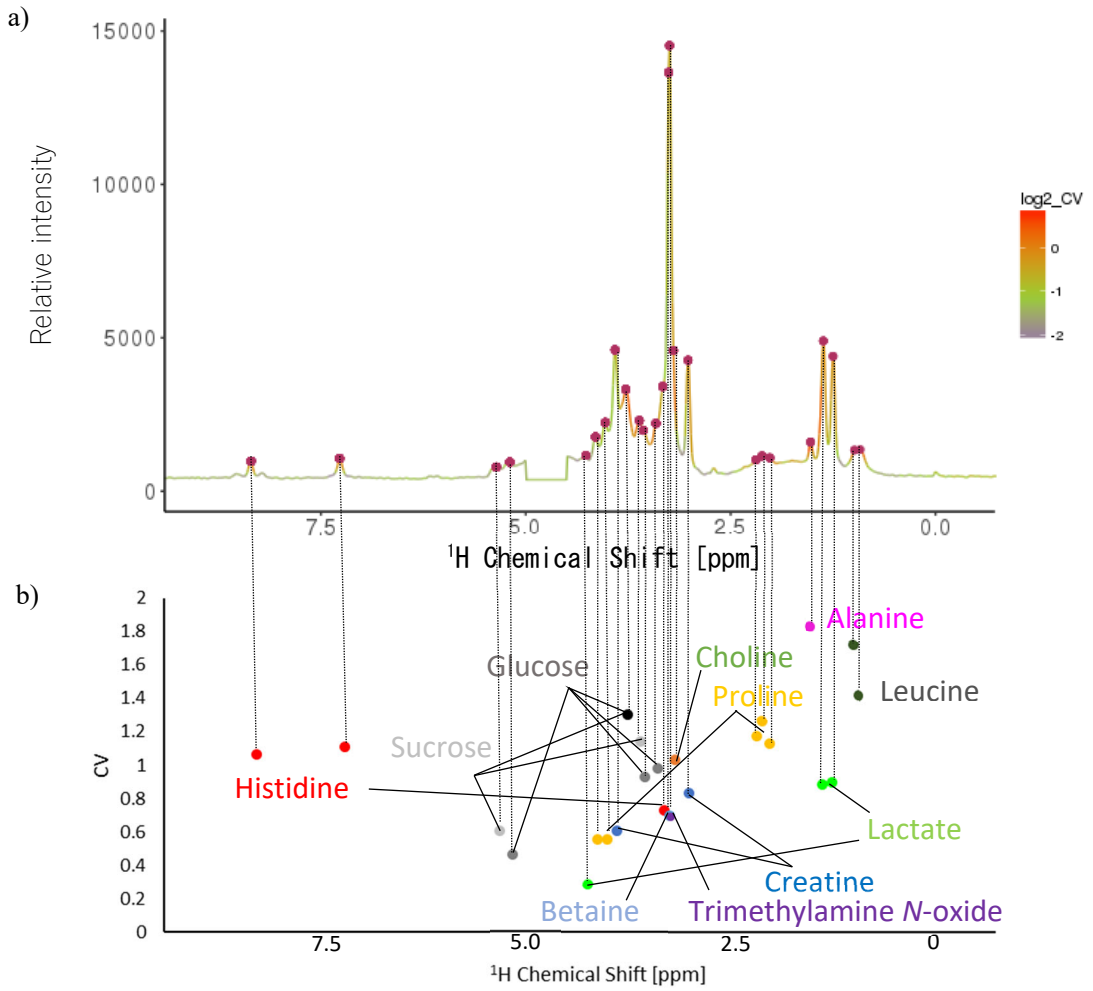
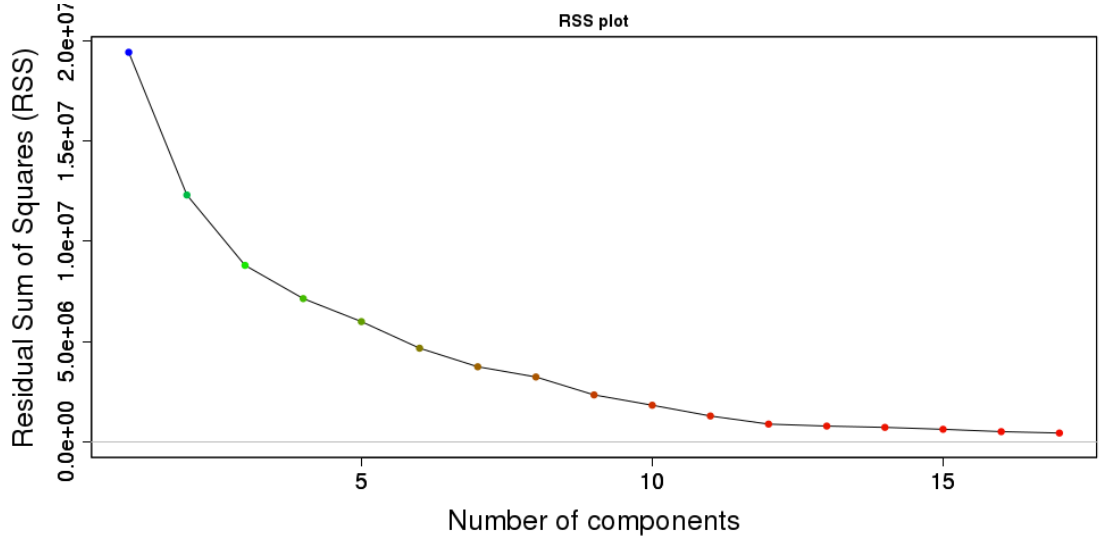


Figure S4. SENSI (SENsitivity improvement with Spectral Integration) result for integrated 51 fish spectra of fish extracts obtained by benchtop 60 MHz NMR. (**a**) Picked peaks, and (**b**) CV.

a)



b)

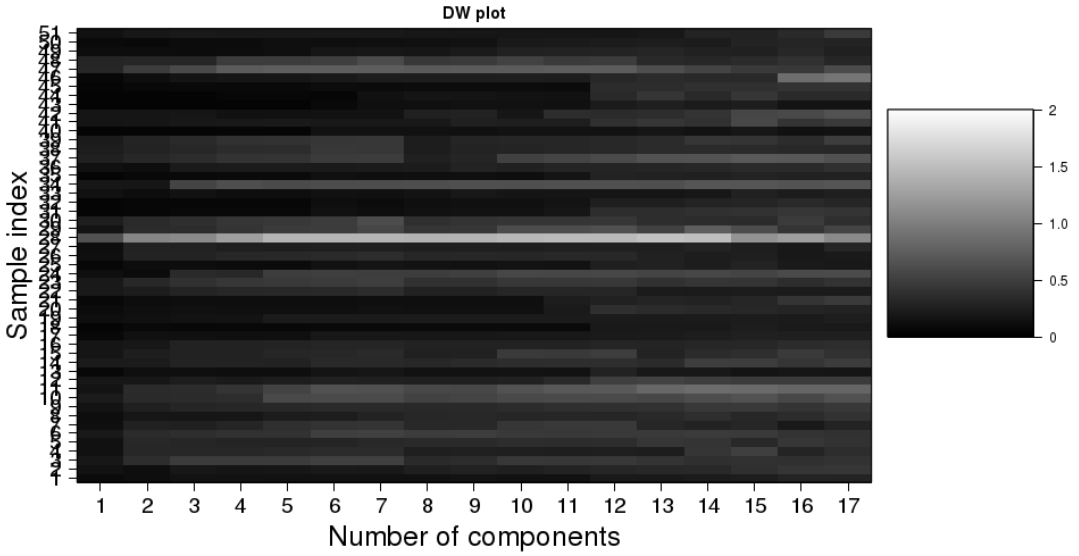


Figure S5. PKSP results of 51 fish spectra using the NNSC method. (**a**) RSS plot, and (**b**) DW plot.

c)

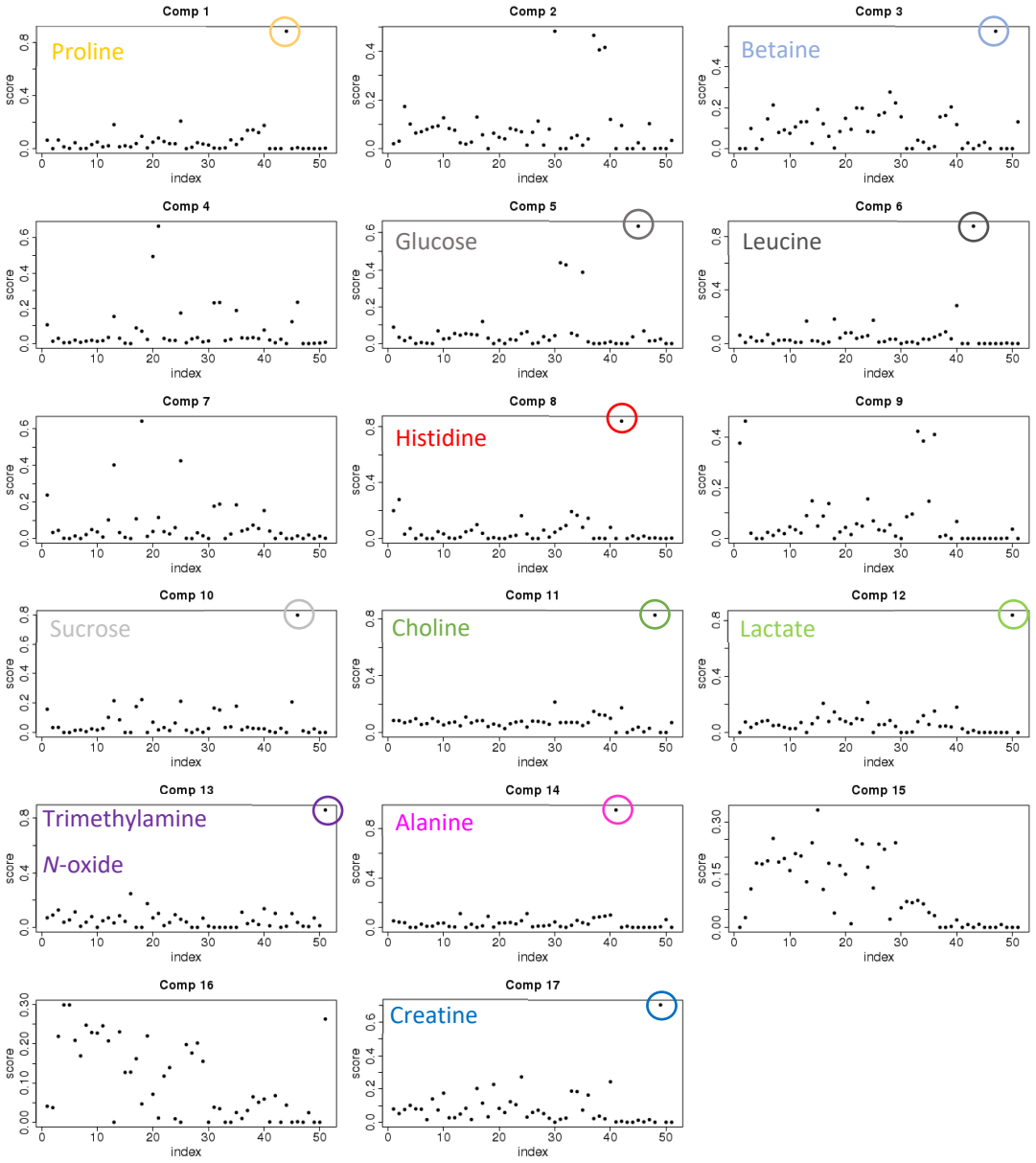


Figure S5. PKSP results of 51 fish spectra using the NNSC method (continuation). (**c**) Plot of scores for different components.

d)

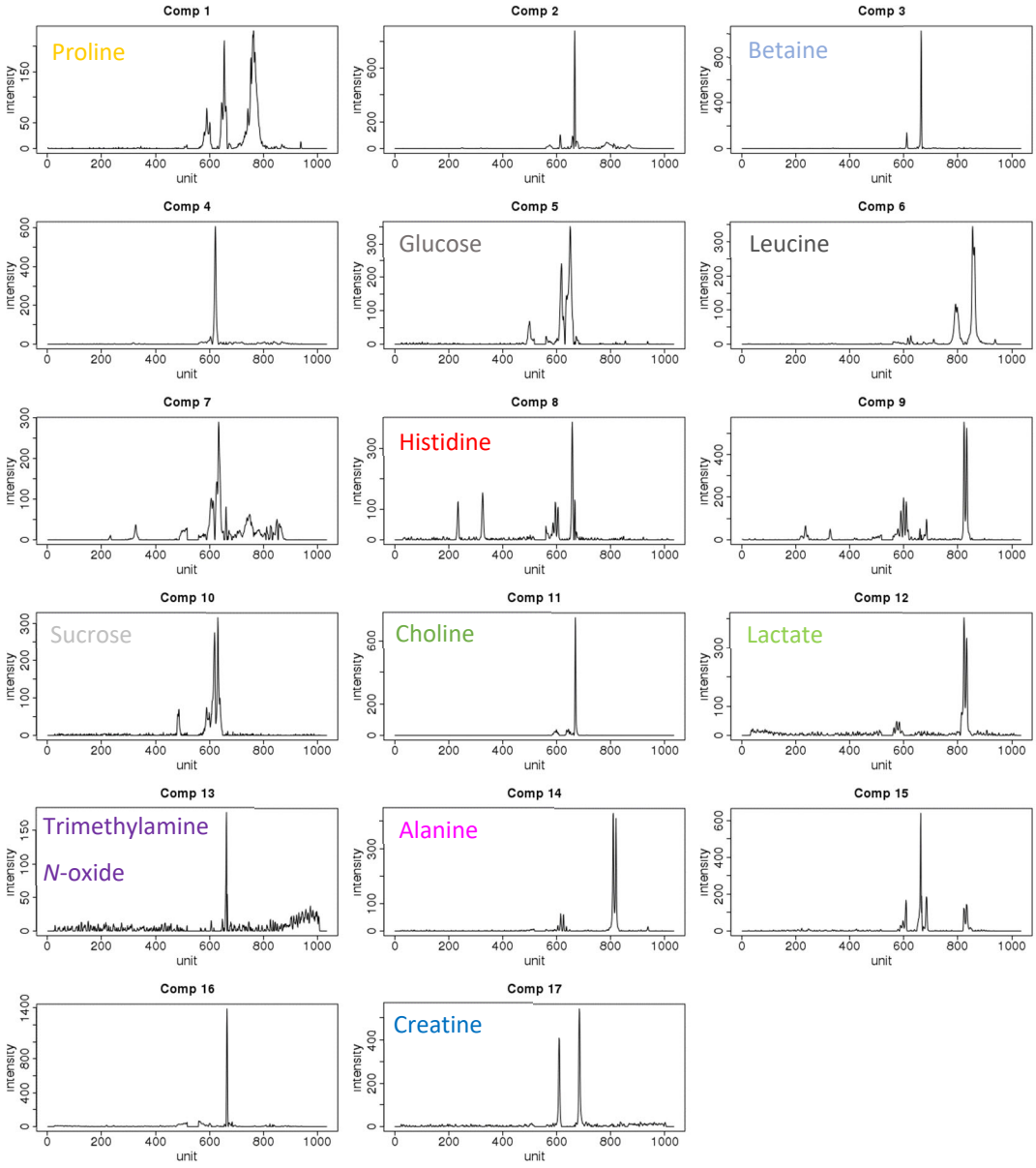


Figure S5. PKSP results of 51 fish spectra using the NNSC method (continuation). (**d**) Plot of loadings for different components.

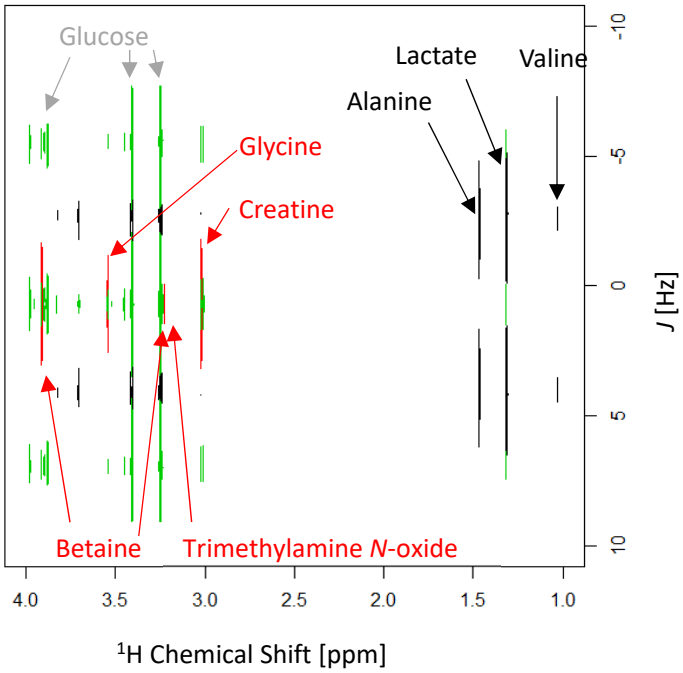


Figure S6. The result of separating one spectrum of 2D-*J*res from *Acanthogobius flavimanus* (Yellowfin goby) body muscle extracts in D_2O by PKSP. As a result of peak separation using PKSP's MCR-ALS algorithm, it was separated into three components of singlet (red), doublet (black) and triplet (multiplet, green). As shown in figure S6, 8 compounds (Valine, Lactate, Alanine, Creatine, Trimethylamine *N*-oxide, Betaine, Glycine and Glucose) were assigned.

Summary

QUERY PEAK	ANNOTATED PEAK
35	31 (88.5/14285/1429%)

	QUERY		ANNOTATION	
PEAK	^{1}H PPM	^{13}C PPM	COMPOUND & ATOM	COUNTS
1		180.7	Lignin(Nakarai) ; Xylan(Birchwood) ; Xantangum ; Lipid ;	4
2		172.7	Chitin ; Peptide ;	2
3		156.8	Peptide ;	1
4		148.3	Lignin(Nakarai) ; Lignin(Nakarai) ;	2
5		136.5	Peptide ;	1
6		129.3	Lignin(Nakarai) ; HumicAcid ; Peptide ; Peptide ; Lipid ;	5
7		128.3	Lignin(Nakarai) ; HumicAcid ; HumicAcid ; Peptide ;	4
8		118.2	Lipid ;	1
9		105.3	AlphaCellulose ; Rhamnogalacturonan ; Inulin ; Carrageenan ; Chitosan ; Galactan(Potate) ; Curdran ; Agarose ; Agarose ; Cellulose ; Paramylon ;	11
10		103.5	Laminarin ; Galactan ; Rhamnogalacturonan ; Arabinoxylan(Wheat) ; Arabinoxylan(Wheat) ; Inulin ; Inulin ; Chitin ; Xyloglucan ; Starch(Wheat) ; Curdran ; Pullulan ; Lichenan ; Lichenan ; Laminaran ; Laminaran ; Agarose ; Agarose ; Peptide ;	19
11		100.9	AlginicAcid ; AlginicAcid ; AlginicAcid ; Fucoidan ; Pectin(Apple) ; Rhamnogalacturonan ; Arabinoxylan(Wheat) ; Xylan(Birchwood) ; Starch(Wheat) ; Starch(Wheat) ; PolygalacturonicAcid ; PolygalacturonicAcid ; Galactomannan(Guar) ; Agarose ; Agarose ;	15
12		92.7	Galactan ; Fucoidan ; Lichenan ; Lichenan ; Dextran ; Dextran ;	6
13		90.4	Laminarin ; Laminarin ; Carrageenan ; Curdran ; Curdran ; Lichenan ; Lichenan ; Laminaran ;	8
14		83.5	Laminarin ; AlphaCellulose ; AlphaCellulose ; Rhamnogalacturonan ; Arabinoxylan(Wheat) ; Inulin ; Chitin ; Xyloglucan ; Chitosan ; Chitosan ; Arabinan ; Pullulan ; Lichenan ; Agarose ; Xantangum ; Cellulose ; Paramylon ;	17
15		79.7	AlginicAcid ; Fucoidan ; Pectin(Apple) ; Pectin(Apple) ; Pectin(Apple) ; Rhamnogalacturonan ; Arabinoxylan(Wheat) ; Galactomannan(Carob) ; Glucomannan ; PolygalacturonicAcid ; PolygalacturonicAcid ; PolygalacturonicAcid ; Galactan(Potate) ; Arabinan ; Arabinan ; Lichenan ; Agarose ; Xantangum ;	18
16		75.9	Laminarin ; AlphaCellulose ; AlginicAcid ; Pectin(Apple) ; Inulin ; Chitin ; Starch(Wheat) ; Chitosan ; Curdran ; Lichenan ; Agarose ; Paramylon ;	12
17		73.1	AlphaCellulose ; Galactan ; Fucoidan ; Chitin ; Lignin(Nakarai) ; Xyloglucan ; Glucomannan ; Galactan(Potate) ; Arabinan ; Arabinan ; Arabinan ; Curdran ; Curdran ; Pullulan ; Xantangum ; Xantangum ; Cellulose ; Peptide ;	18
18		68.8	Laminarin ; AlginicAcid ; Fucoidan ; Pectin(Apple) ; Lignin(Nakarai) ; Carrageenan ; PolygalacturonicAcid ; Galactan(Potate) ; Arabinan ; Xantangum ; Paramylon ;	11
19		62.6	Laminarin ; Galactan ; Galactan ; Xylan(Beechwood) ; Arabinoxylan(Wheat) ; Inulin ; Xylan(Birchwood) ; Galactomannan(Carob) ; Galactomannan(Carob) ; Xyloglucan ; Glucomannan ; Carrageenan ; Starch(Wheat) ; Galactan(Potate) ; Arabinan ; Curdran ; Pullulan ; Pullulan ; Pullulan ; Galactomannan(Guar) ; Laminaran ; Laminaran ; Agarose ; Cellulose ; Paramylon ;	25
20		59.4	Galactan ; Galactan ; Lignin(Nakarai) ; Galactan(Potate) ; Arabinan ; Agarose ;	6
21		53.1	Pectin(Apple) ; Peptide ;	2
22		48.3		
23		42.8	Xylan(Beechwood) ;	1
24		39.6	Peptide ;	1
25		37.4		
26		33.7	Peptide ; Lipid ;	2
27		32.3	Galactan ; Fucoidan ; Peptide ; Lipid ;	4
28		30.1	Lignin(Nakarai) ; HumicAcid ; Peptide ; Lipid ;	4
29		27.3		
30		24.8	Arabinoxylan(Wheat) ; Lignin(Nakarai) ; Xylan(Birchwood) ; Xantangum ; Xantangum ; Peptide ; Lipid ;	7
31		24.3	Arabinoxylan(Wheat) ; Lignin(Nakarai) ; Xylan(Birchwood) ; Xantangum ; Peptide ; Peptide ; Lipid ;	7
32		22.9	Chitin ; Peptide ;	2
33		20.5	Fucoidan ; Lignin(Nakarai) ; HumicAcid ; Xantangum ; Peptide ;	5
34		14.3	Lignin(Nakarai) ; Peptide ; Lipid ;	3
35		12.3		

QUERY PEAK=Query peak No., QUERY 1H PPM=1H chemical shift of query, QUERY 13C PPM=13C chemical shift of query, ANNOTATION COMPOUND & ATOM=All annotated compounds with atom names, ANNOTATION COUNTS=No. of annotations

Figure S7. SpinMacro assignment results of peaks picked by SENSI of the previously reported solid-state CP-MAS spectra of *Euglena gracilis* and standards (paramylon, peptides, lipids). Regarding the assigned chemical shift, paramylon is indicated by the red frame, peptides by the black frame, and lipids by the blue frame. The ^{13}C tolerance is 1 ppm.

a)

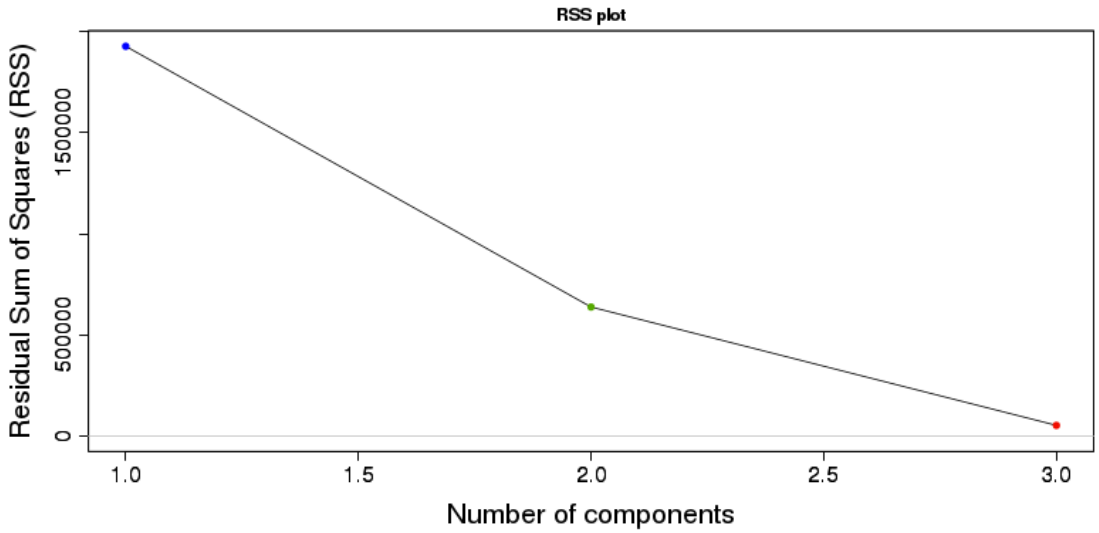


Figure S8. Signal separattion by NNSC of PKSP from *E. gracilis* CP-MAS spectrum. (**a**) RSS plot shows that there are three main components.

b)

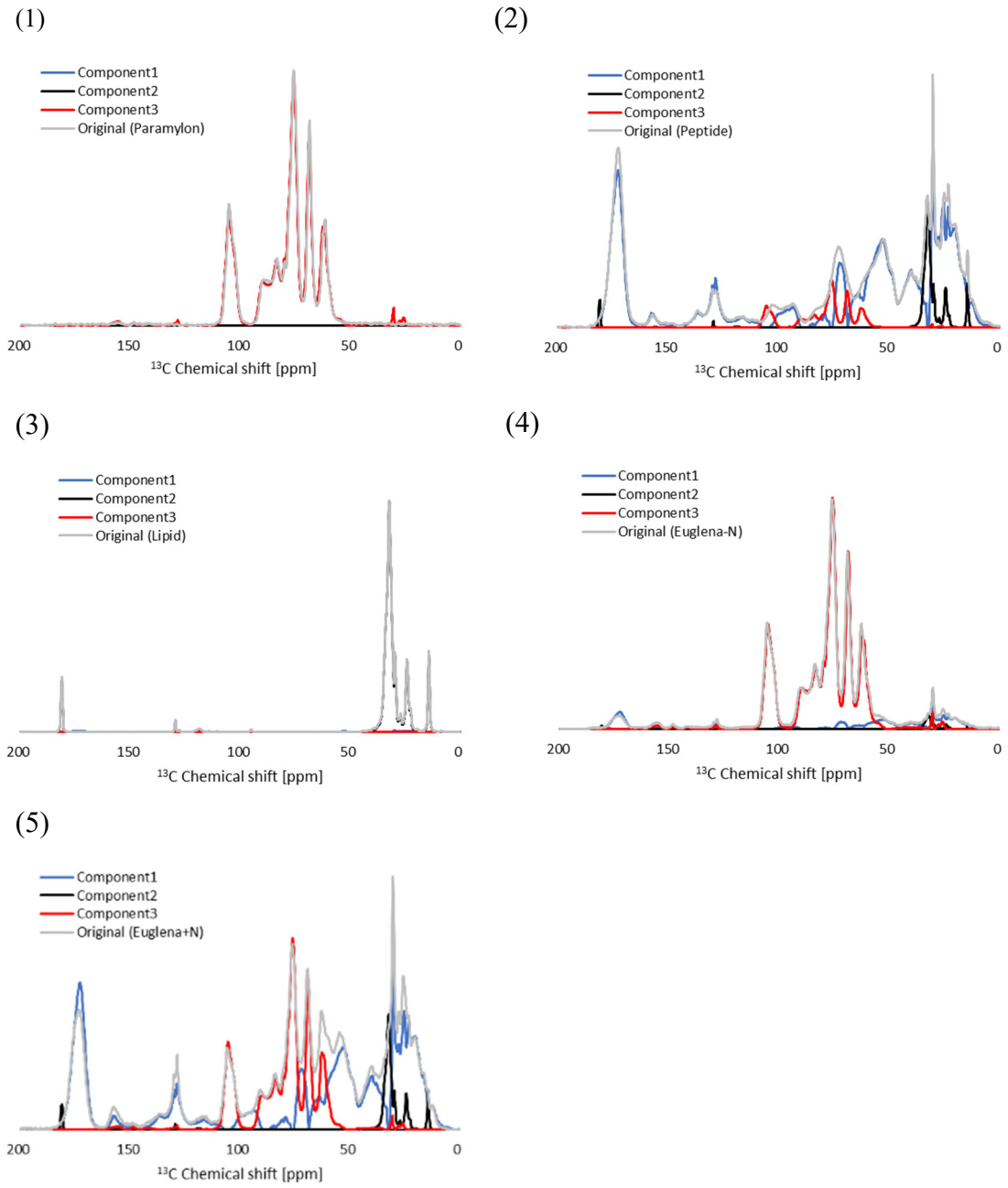


Figure S8. Signal separation by NNSC of PKSP from *E. gracilis* CP-MAS spectrum (continuation). (**b**) Separated spectra of paramylon, peptides, and lipids. Component 1 was identified as peptides (b-1), component 2 was identified as lipids (b-2), and component 3 was identified as paramylon (b-3). The *E. gracilis* sample was separated into 3 components (b-4, b-5). The variation in paramylon and peptides components indicates metabolic fluctuation of *E. gracilis*.

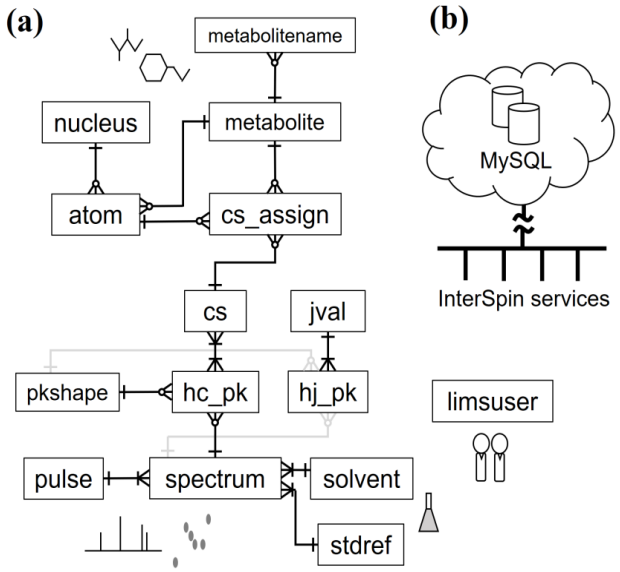


Figure S9. Schematic diagram of SpinLIMS.

(**a**) SpinLIMS is a relational database based on NMR spectra and molecular information. Thirty-four tables store a variety of information: for example, “metabolitename” is compound name, “metabolite” is compound, “atom” is atom, “nucleus” is nuclide, “cs_assign” is assignment of chemical shift, “cs” is chemical shift, “Jval” is *J* values, “hc_pk” is HSQC peaks (“h_pk” for ^{1}H-1D NMR, “c_pk” for ^{13}C-1D NMR), “pkshape” is peak linear type, “hj_pk” is 2D-*J*res peak, “spectrum” is spectrum, “pulse” is NMR pulse sequence type, “solvent” is solvent (register as “none” in case of solid), “stdref” is reference compound. (**b**) SpinLIMS services are provided to users from the MySQL server via InterSpin (http://dmar.riken.jp/interspin/).

a)

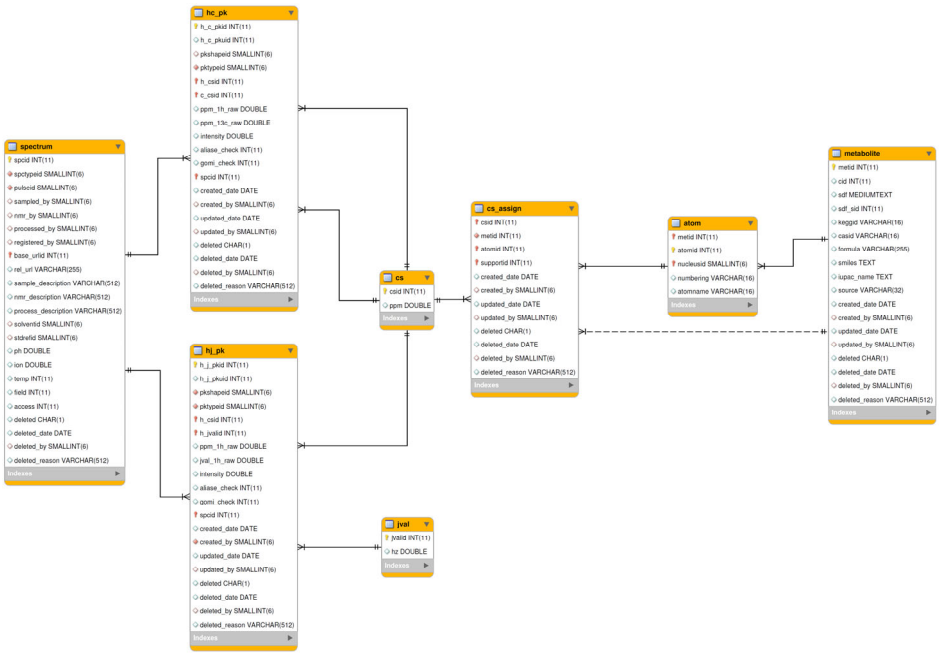


Figure S10. Diagram showing the relationship among different entities on the SpinLIMS database. (**a**) Core entities or "tables" within the relational database.

b)

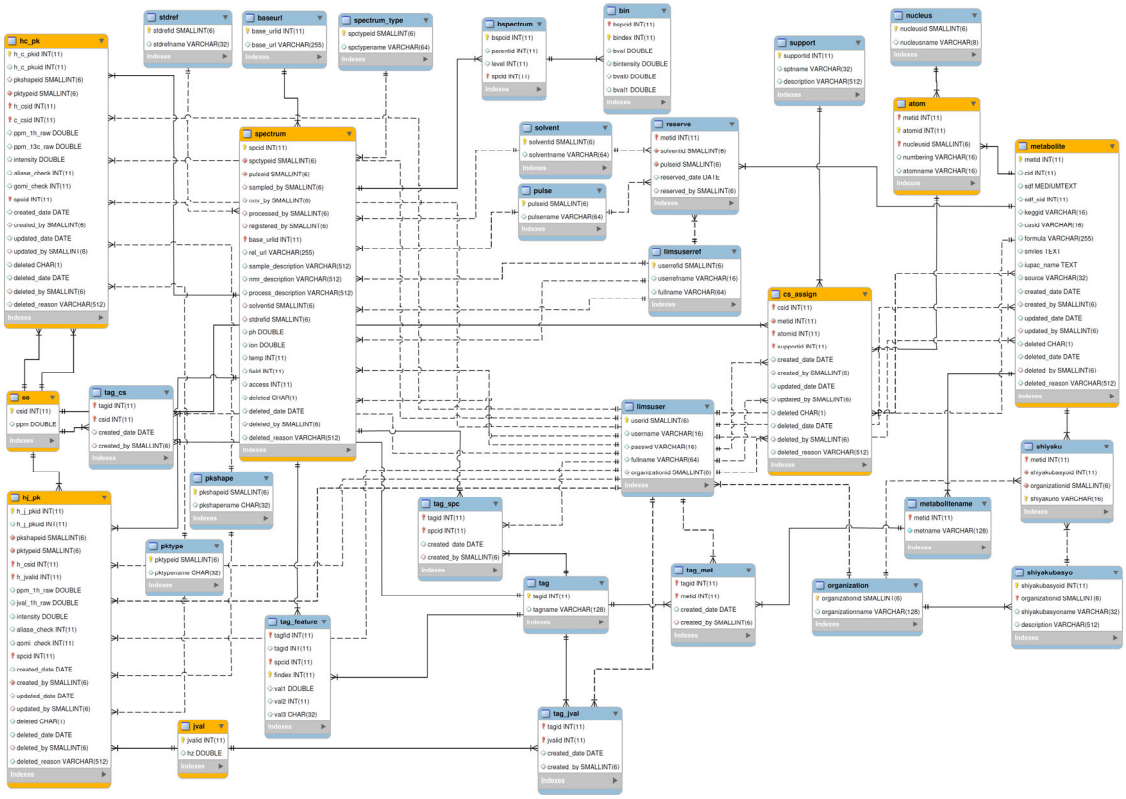


Figure S10. Diagram showing the relationship among different entities on the SpinLIMS database (continuation). (**b**) The complete relationship among all entities (tables). For a given query, each table contains the following data: “hc_pk” is HSQC peaks; “hj_pk” is 2D-*J*res peaks; “h_pk” is ^{1}H-1D NMR peaks; “c_pk” is ^{13}C-1D NMR peaks; “spectrum” is the spectrum; “spectrum_type” is the type of spectrum; “pulse” is the NMR pulse sequence; “solvent” is the solvent; “stdref” is the standard compound; “baseurl” is the base URL for raw spectrum; “pkshape” is the peak shape; “pktype” is the type of peak; “cs” is chemical shifts; “jval” is *J* values; “cs_assign” is one-atom assignment for one chemical shift; “support” is evidence of assignment; “metabolite” is metabolite; “atom” is atom; “nucleus” is nucleus; “metabolitename” is the name of metabolite; “shiyaku” is the compound number; “Shiyakubasyo” is the storage location of compound; “organization” is the organization who measured the data; “limsuser” is the user who registered the data; “reserve” is the measuring reservation; “tag” is the tag for clustering compound, metabolite, and spectrum; “tag_met” is the tag for metabolite; “tag_cs” is the tag for chemical shift; “tag_jval” is the tag for *J* value; “limsuserref” is the reference user for sampled_by, nmr_by, processed_by of spectrum table, and reserved_by of reserve table; “tag_feature” is a feature to tag related spectra, for example, for tasting data, etc.; “bin” is the bin for spectrum; “bspectrum” is the binning spectrum; and “tag_spc” is the tag for spectrum. The complete entity relationship diagram is published at http://dmar.riken.jp/interspin/

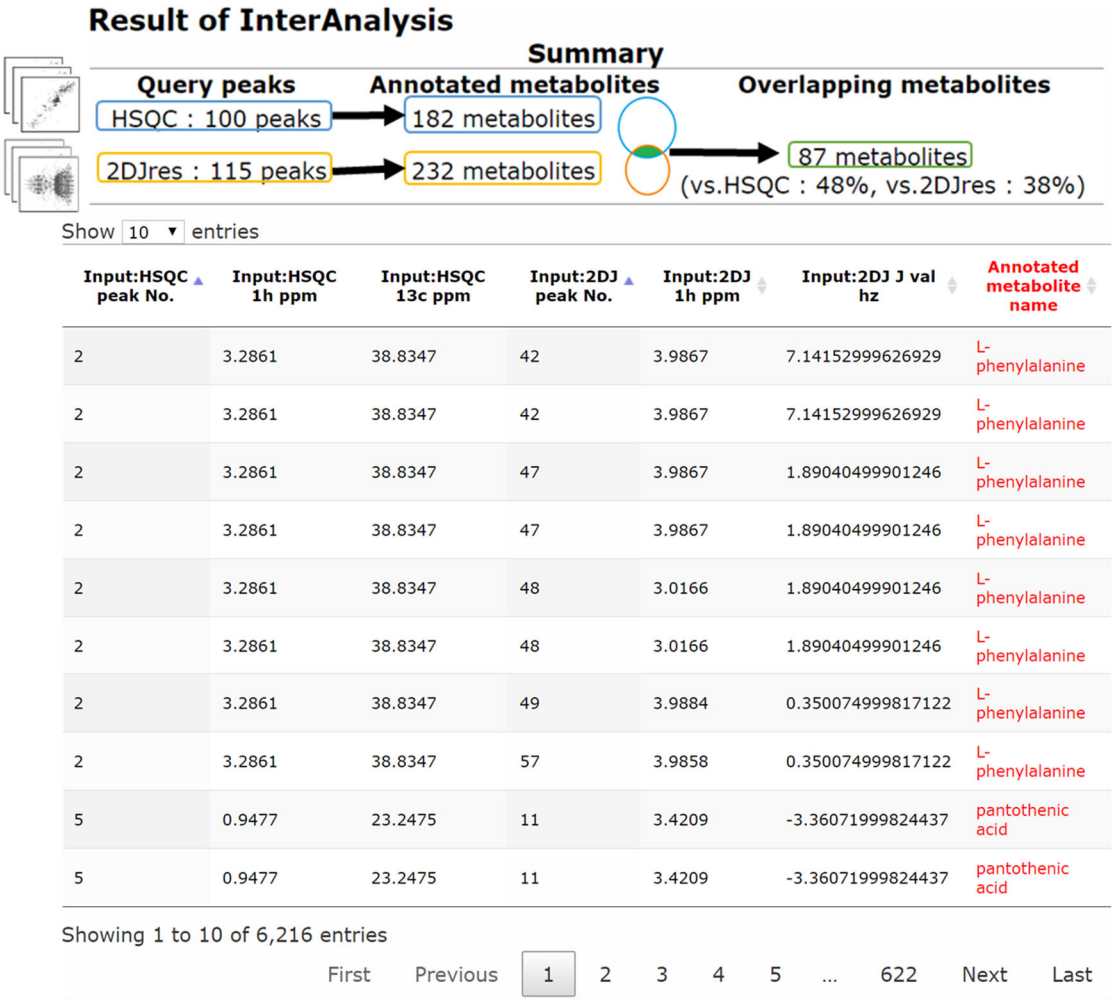


Figure S11. Result of InterAnalysis for HSQC and 2D *J*res peaks from *Acanthogobius flavimanus* body muscle extracted in deuterated potassium phosphate (KPi). The summary shows the number of query peaks, the number of assigned molecules, and the narrowed down set of molecules. The table shows some of the molecular assignment results for each query peak. For data acquired in KPi extract, the original webtools SpinAssign and SpinCouple assigned 182 and 232 molecules, respectively. By contrast, InterAnalysis assigned 87 molecules, thereby narrowing the down molecules to 48% and 38%, respectively.

Table S1. List of compounds in mixtures 1 and 2.

No.	Compound	Mixture 1 (mM)	Mixture 2 (mM)
1	Alanine	5	2
2	Leucine	5	5
3	Lysine	2	2
4	Phenylalanine	3	5
5	Proline	5	5
6	Threonine	3	5
7	Valine	2	2
8	Malate	5	10
9	Glucose	10	3
10	Sucrose	5	5

Table S2. List of 40 fish samples.

No.	Scientific name	Recipe	Storage time
1	*Auxis rochei rochei*	Soy marinated, Fried	—
2	*Auxis rochei rochei*	Grilled	—
3	*Branchiostegus japonicus*	Fried	—
4	*Epinephelus areolatus*	Fried	1day
5	*Epinephelus areolatus*	Fried	2day
6	*Epinephelus septemfasciatus*	Raw	—
7	*Lateolabrax japonicus*	Raw	—
8	*Lateolabrax japonicus*	Raw	—
9	*Lateolabrax japonicus*	Steamed rice wine	—
10	*Lateolabrax japonicus*	Steaming	—
11	*Nemipterus virgatus*	Fried	—
12	*Nemipterus virgatus*	Mayonnaise	—
13	*Pagrus major*	Soy marinated	1day
14	*Pagrus major*	Fried	—
15	*Pagrus major*	Grilled	—
16	*Pagrus major*	Raw	—
17	*Pagrus major*	Mayonnaise	—
18	*Paralichthys olivaceus*	Soy marinated, Fried	1day
19	*Paralichthys olivaceus*	Raw	—
20	*Paralichthys olivaceus*	Marinated kelp	1hour
21	*Paralichthys olivaceus*	Kelp, Salt, Citrus depressa	4day
22	*Paralichthys olivaceus*	Raw	—
23	*Paralichthys olivaceus*	Yuzu, Salt, Grilled	—
24	*Sardinops melanostictus*	Raw	—
25	*Scombrops gilberti*	Soy marinated	—
26	*Scombrops gilberti*	Raw	—
27	*Scombrops gilberti*	Steaming	2day
28	*Scorpaenopsis cirrhosa*	Fried	—
29	*Scorpaenopsis cirrosa*	Raw	—
30	*Sepia esculenta*	Raw	—
31	*Seriola quinqueradiata*	Grilled miso	1hour
32	*Seriola quinqueradiata*	Grilled miso	2hour
33	*Thunnus*	Raw	—
34	*Thunnus*	Raw	1day
35	*Thunnus orientalis*	Soy marinated	—
36	*Thunnus orientalis*	Grilled	—
37	*Todarodes pacificus*	Dried	—
38	*Todarodes pacificus*	Dried	1day
39	*Todarodes pacificus*	Dried	3day
40	*Zeus faber*	Raw	—

Table S3. List of standard compounds.

No.	Compound
41	Alanine
42	Histidine
43	Leucine
44	Proline
45	Glucose
46	Sucrose
47	Betaine
48	Choline
49	Creatine
50	Lactate
51	Trimethylamine *N*-oxide

Table S4. Improvement of signal-to-noise ratio by SENSI.

^{1}H chemical shift of picked peaks [ppm]	**Intensity of picked peaks by SENSI**	**Average intensity of original spectrum**	**S/N after SENSI**	**S/N before SENSI**	**Sensitivity improvement rate**	**Peaks enhancement rate**
8.35	980.1	12.3	16.0	2.4	6.7	79.9
7.27	1063.5	13.9	17.3	2.7	6.5	76.5
5.36	788.8	8.5	12.9	1.6	7.8	92.6
5.19	947.8	11.6	15.4	2.2	6.9	81.4
4.27	1162.9	15.9	19.0	3.1	6.2	73.3
4.15	1770.3	27.8	28.8	5.4	5.4	63.8
4.03	2241.1	37.0	36.5	7.1	5.1	60.6
3.91	4606.0	83.4	75.1	16.1	4.7	55.2
3.78	3318.3	58.1	54.1	11.2	4.8	57.1
3.61	2302.3	38.2	37.5	7.4	5.1	60.3
3.57	1979.2	31.9	32.3	6.1	5.3	62.1
3.41	2211.9	36.4	36.0	7.0	5.1	60.7
3.33	3409.6	59.9	55.6	11.5	4.8	56.9
3.26	13649.8	260.7	222.4	50.3	4.4	52.4
3.25	14525.5	277.9	236.7	53.6	4.4	52.3
3.20	4585.9	83.0	74.7	16.0	4.7	55.3
3.02	4268.6	76.8	69.6	14.8	4.7	55.6
2.19	1028.2	13.2	16.8	2.5	6.6	77.8
2.12	1149.3	15.6	18.7	3.0	6.2	73.7
2.02	1087.3	14.4	17.7	2.8	6.4	75.6
1.52	1602.3	24.5	26.1	4.7	5.5	65.5
1.37	4897.7	89.1	79.8	17.2	4.6	55.0
1.25	4397.1	79.3	71.7	15.3	4.7	55.5
0.99	1340.2	19.3	21.8	3.7	5.9	69.3
0.92	1357.9	19.7	22.1	3.8	5.8	69.0
Average	3226.9	56.3	52.6	10.9	5.5	65.5

Signal-to-noise (max–min, 9–10 ppm) is 5.2 before SENSI and 61.4 after SENSI.

Appendix B

Supplementary material for

Signal Deconvolution and Noise Factor Analysis based on a Combination of Time–Frequency Analysis and Probabilistic Sparse Matrix Factorization

Supplementary material

1. The mathematical theory of signal deconvolution
2. Supplementary figures and tables
3. References

An NMR measurement informatics tool is available at http://dmar.riken.jp/NMRinformatics/.

1. The Mathematical Theory of Signal Deconvolution

In this study, signal deconvolution was applied to free induction decays (FIDs) of one-dimensional (1D) nuclear magnetic resonance (NMR) to separate components and improve the signal-to-noise ratio (SNR). The mathematical theory underlying this signal deconvolution is based on the combined methods of short-time Fourier transform (STFT) and probabilistic sparse matrix factorization (PSMF).

In Fourier transform (FT) NMR spectroscopy, an FID is the NMR signal generated by non-equilibrium nuclear spin magnetization precessing along the magnetic field. This non-equilibrium magnetization can be generated by applying a pulse of resonant radiofrequency close to the Larmor frequency of the nuclear spins in the sample. An FID is usually a sum of multiple decayed oscillatory signals. These signals return to equilibrium at different rates or relaxation time constants. Analysis of the relaxation times of an FID for a sample gives significant insight into the chemical composition, structure, and mobility of that sample. FIDs acquired by NMR measurement are composed of many signals derived from the sample and several types of noise, such as external noise, physical vibration, power supply, and internal noise of the spectrometer due to thermal noise. Therefore, an FID signal can be modeled as:

$$S(t) = S_{signal}(t) + S_{noise}(t). \quad \text{(S1)}$$

where $S(t)$ is the measured signal, and $S_{\text{signal}}(t)$ and $S_{\text{noise}}(t)$ represent a set of ideal signals and a set of signals from different types of noise (Equation (S1)) [1]. Suppose that a 90° pulse is applied to an equilibrium magnetization along the *z*-axis, resulting in magnetization of the *x*–*y* plane, which then precesses in the transverse plane with angular frequency Ω. The corresponding time-domain signal that decays with time *t* is the FID *S*(*t*). In principle, the exponential decay constant of the FID is the T_2 relaxation time, which is a physically parameter independent of field inhomogeneity. In reality, however, because of the effect of magnetic field homogeneity, the decay constant of the FID is called T_2*, an instrument-dependent parameter, rather than T_2. *S*(*t*) is given by the relaxation time constant T_2^* [2]:

$$S(t) = S_0 \exp(i\Omega t) \exp\left(-\frac{t}{T_2^*}\right), \quad \text{(S2)}$$

where S_0 is the initial transverse magnetization at $t = 0$ immediately after the 90° pulse (Equation (S2)). The relaxation process can be described by saying that the transverse magnetization $S(t)$ decays exponentially according to Equation (S2). The shorter the relaxation time T_2^*, the more rapid the decay.

If an FID has more than one component, the FID will be the sum of contributions from each component (Equation (S3)):

$$S(t) = \sum_{k=1}^{n} S_{0k} \exp(i\Omega_k t) \exp\left(-\frac{t}{T_{2k}^*}\right). \quad \text{(S3)}$$

When there are two or more types of component (i.e., $k > 1$) in the FID signal, it is difficult to determine the individual signals from the time-domain signal $S(t)$. Therefore, we apply FT to $S(t)$ to yield a frequency-domain spectrum $S(\omega)$ with an angular frequency variable ω on the horizontal axis and *k* peaks at Ω_k (Equation (S4)):

$$S(\omega) = \int_{-\infty}^{\infty} S(t)\exp(-i\omega t)dt. \quad \text{(S4)}$$

Standard FT (Equation (S4)) has only has the frequency domain; therefore, we apply STFT, which has both frequency and time domains. Because the FID signal decays exponentially with time, for STFT, it needs to be divided into several small time intervals (i.e., segments) to analyze the time–frequency feature accurately, and FT is used to determine the frequency feature of each segment, thereby increasing the accuracy of signal feature extraction. STFT uses a window function to obtain each weighted segment on the time axis and then applies FT to the segment. STFT of $S(t)$ can be written as:

$$STFT_S(\tau, \omega) = \int_{-\infty}^{\infty} S(t)g(t-\tau)\exp(-i\omega t)dt, \quad \text{(S5)}$$

where the window function g is first used to intercept the progress of FT on $S(t)$ around $t = \tau$ locally, and then FT of the segment is performed on *t* (Equation (S5)) [3]. By moving the center position of the window function g sequentially, all of the FTs at different times can be obtained.

Applying Euler's formula (Equation (S6)),

$$\exp(-i\omega t) = \cos \omega t - i \sin \omega t, \quad \text{(S6)}$$

shows that the value of $STFT_S(\tau, \omega)$ is complex and composed of two signals, a real part (Re) and an imaginary part (Im), whose phases differ by 90° from each other (Figure S1, Equation (S7) and (S8)):

$$Re = \gamma \cos \omega\tau, \quad \text{(S7)}$$

$$Im = \gamma \sin \omega\tau\,. \quad \text{(S8)}$$

To change a complex value into an absolute value, the following equation is applied (Equation (S9)):

$$|z| = \sqrt{Re^2 + Im^2}. \quad \text{(S9)}$$

For PSMF [4], positive-valued matrices are needed and the original signal values must be converted to their logarithmic form for optimal analysis. To convert Equation (S9) to a positive logarithmic form, the following equation is applied (Equation (S10)):

$$V = \log_{10}(|z| + 1). \quad \text{(S10)}$$

In our method using PSMF, we focus on sparse factorizations and on properly accounting for uncertainties while computing the factorization. Thus, signal deconvolution is formulated as finding the factorization of the data matrix V (Equation (S11)):

$$V = W \cdot H + residuals. \quad \text{(S11)}$$

When considering the separation of signal and noise, Equation (S11) can be described as the sum of a signal component, a noise component, and residuals (Equation (S12)):

$$V = W_{signal} \cdot H_{signal} + W_{noise} \cdot H_{noise} + residuals. \quad \text{(S12)}$$

Equation (S12) estimates that the signal component ($W_{signal} \cdot H_{signal}$) decays exponentially with time, while the noise component ($W_{noise} \cdot H_{noise}$) is a random or flat value. To reconstruct the FIDs, the absolute value within each component is converted back to a complex value using the following equations (Equation (S13) and (S14)):

$$Re = (10^{\log_{10}|z+1|} - 1)\cos\theta, \quad \text{(S13)}$$

$$Im = (10^{\log_{10}|z+1|} - 1)\sin\theta\,. \quad \text{(S14)}$$

The inverse short-time Fourier transform (ISTFT), $S_{inv}(t)$, is computed by overlap-adding the inverse fast Fourier transform signals in each segment of the STFT spectrogram as follows (Equation (S15)) [5]:

$$S_{inv}(t) = \int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} V(\omega)\exp(i\omega t)d\omega. \quad \text{(S15)}$$

To evaluate SNR, both noise-removed and noise-only FIDs are converted to signal and noise spectra, respectively, by applying standard FT. SNR is calculated as the ratio of the signal peak intensity to the noise value by using the method of Mnova (Equation (S16)) [6]:

$$SNR = \frac{Signal\ peak\ intensity}{Noise\ value}. \quad \text{(S16)}$$

The noise value is calculated by using the standard deviation of the signals-free region (Equation (S17)):

$$Noise\ value = \sqrt{\frac{\sum_{i=1}^{N}(S(t)_i - S(t)_m)^2}{N-1}}, \quad \text{(S17)}$$

where N is number of points in the signal-free region, $S(t)_i$ is the value of each digital point in that region, and $S(t)_m$ is average of the digital points in that region.
Finally, the relative SNR is the ratio of the SNR after denoising ($SNR_{denoised}$) to the original SNR ($SNR_{original}$), which is calculated as follows (Equation (S18)):

$$Relative\ SNR = \frac{SNR_{denoised}}{SNR_{original}}. \quad \text{(S18)}$$

In order to obtain a theoretical SNR index based on acquisition parameters, the theoretical SNR value (*calcSNR*) was calculated by using a previously described formula (Equation (S19)) [7]:

$$calcSNR = \frac{C\gamma_{exc}T_2(\gamma_{det}B)^{3/2}\sqrt{NS}}{TE} \propto \frac{C(B)^{3/2}\sqrt{NS}}{TE\nu_{1/2}}. \quad \text{(S19)}$$

where, *C* is the number of spins in the system (sample concentration/number of protons), γ_{exc} is the gyromagnetic ratio of the excited nucleus, γ_{det} is the gyromagnetic ratio of the detected nucleus, *NS* is the number of scans, *B* is the external magnetic field, T_2 is the transverse relaxation time (the reciprocal of π times the line width at half height), *TE* is the sample temperature, and $\nu_{1/2}$ is the full width at half maximum.

2. Supplementary figures and tables

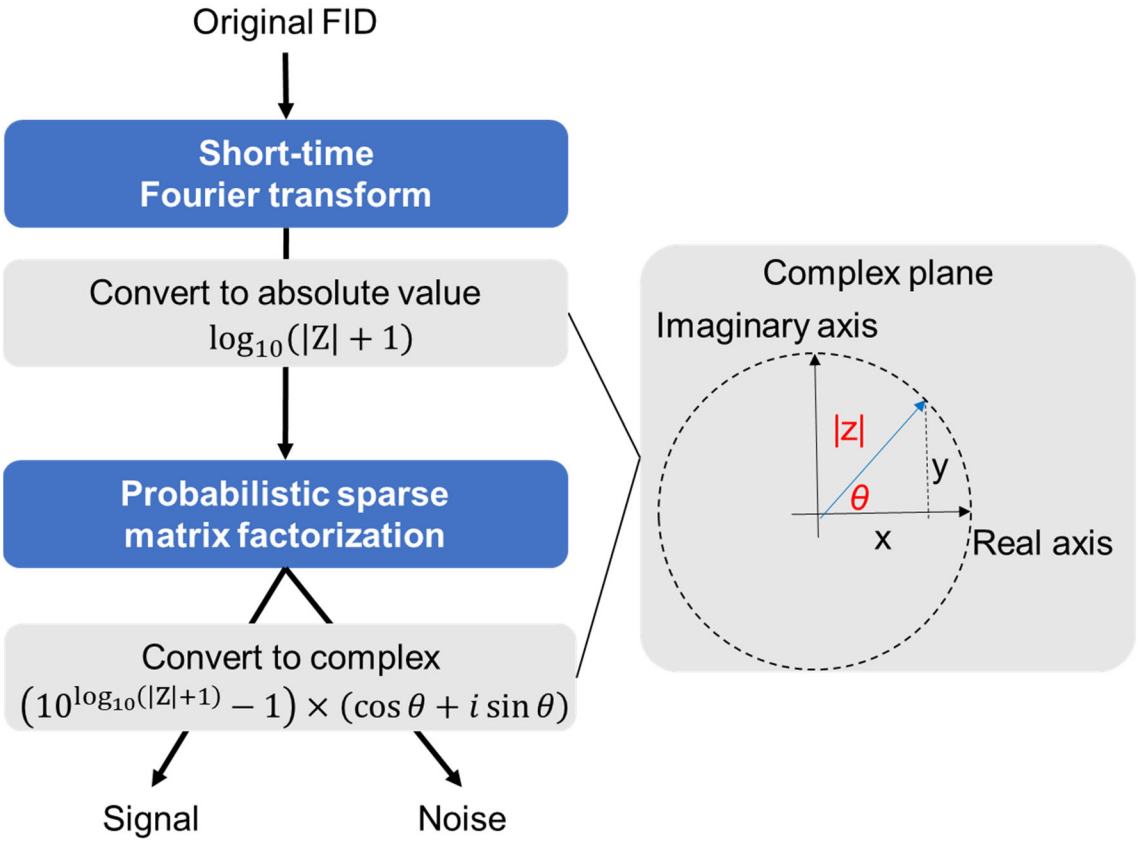


Figure S1. Schematic diagram showing the steps in the signal deconvolution method, including absolute value conversion and complex value conversion of the matrix. The original FID is subjected to STFT. The matrix of STFT is converted to an absolute value. This nonnegative value is separated to components of signal and noise by PMSF. The separated components are then converted to a complex value, from which denoised FIDs and time-domain noise data are extracted. The right image shows the relationship among the real part, imaginary part, absolute value, and argument in the complex plane.

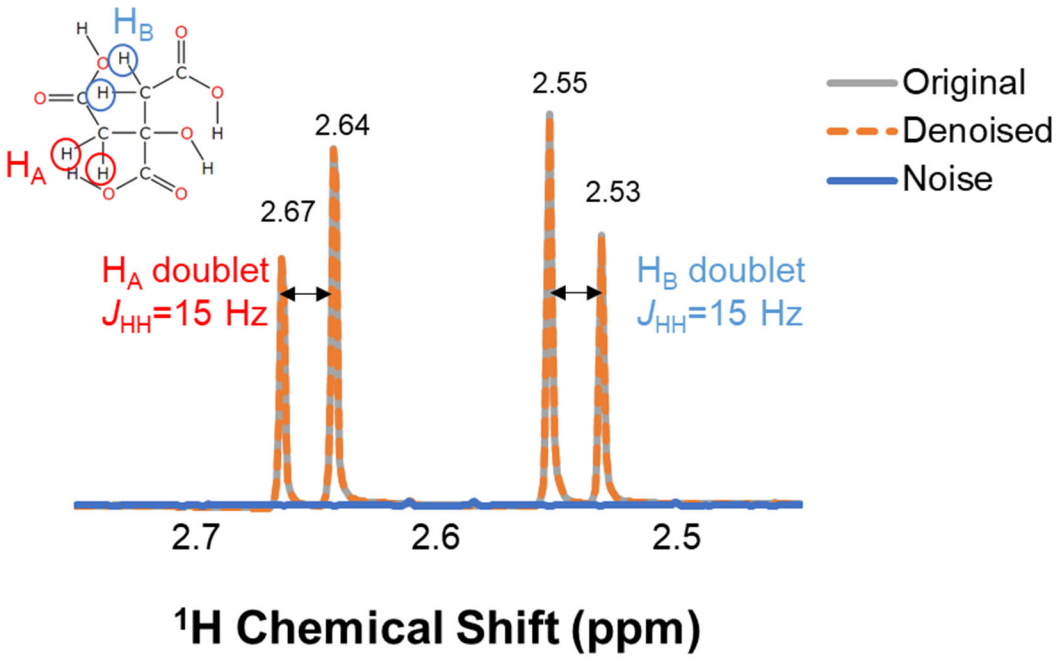


Figure S2. Original spectra and denoised spectra in ^{1}H-NMR data of citric acid. To demonstrate the denoising method, data for citric acid were acquired by using the pre-saturation (program name; "zgpr ") pulse sequence. The original spectrum (grey, solid line), denoised spectrum (orange, dashed line) and noise (blue, solid line) are shown. The chemical structure, peaks and *J* value of citric acid are shown in the figure. Information on the spectral values is shown in Table S1. Relative SNR of this spectra is 1.14-fold.

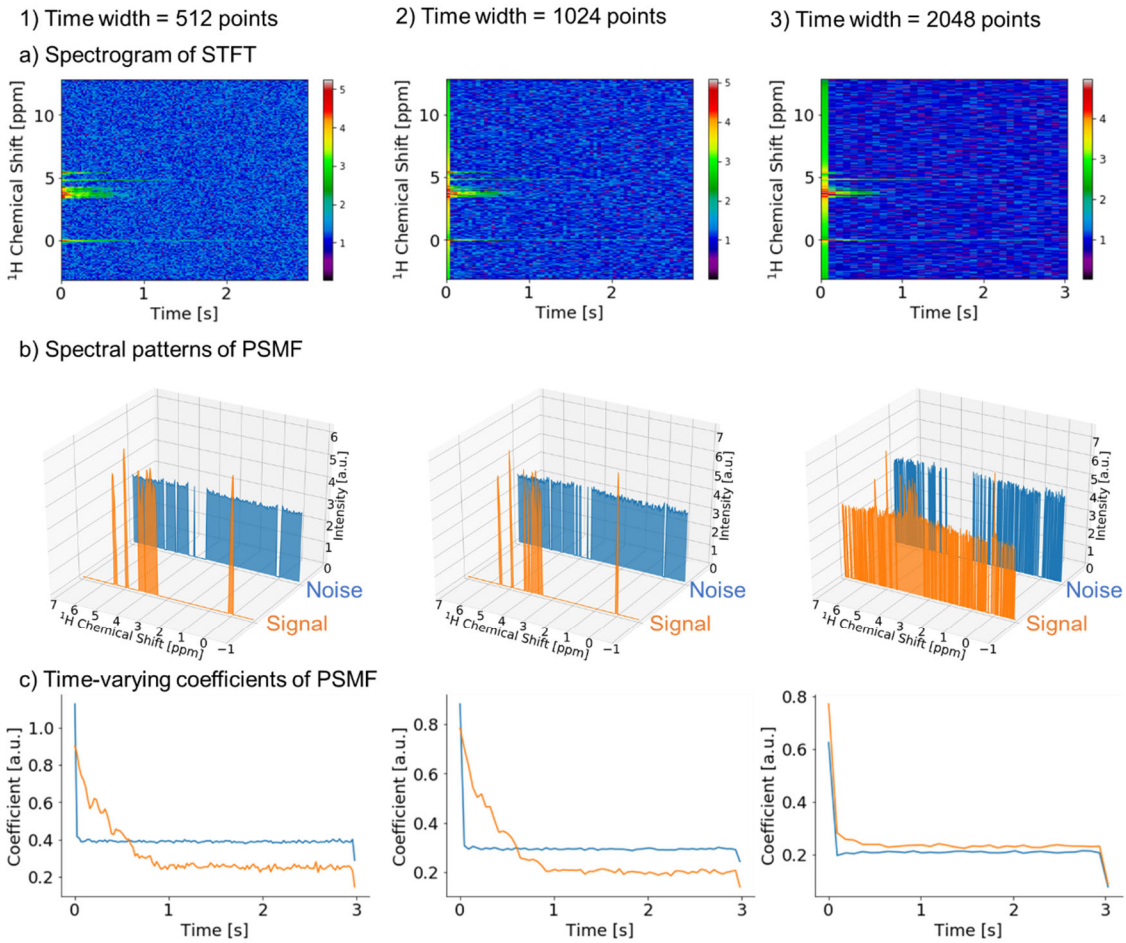


Figure S3. Effect of STFT time width on PSMF. STFT was performed using three different time widths, 512 points (1), 1024 points (2), and 2048 points (3), and the effect on separated components was investigated. a) Spectrogram obtained by STFT. b) Spectral patterns of PSMF. c) Time-varying coefficient of each component in PSMF.

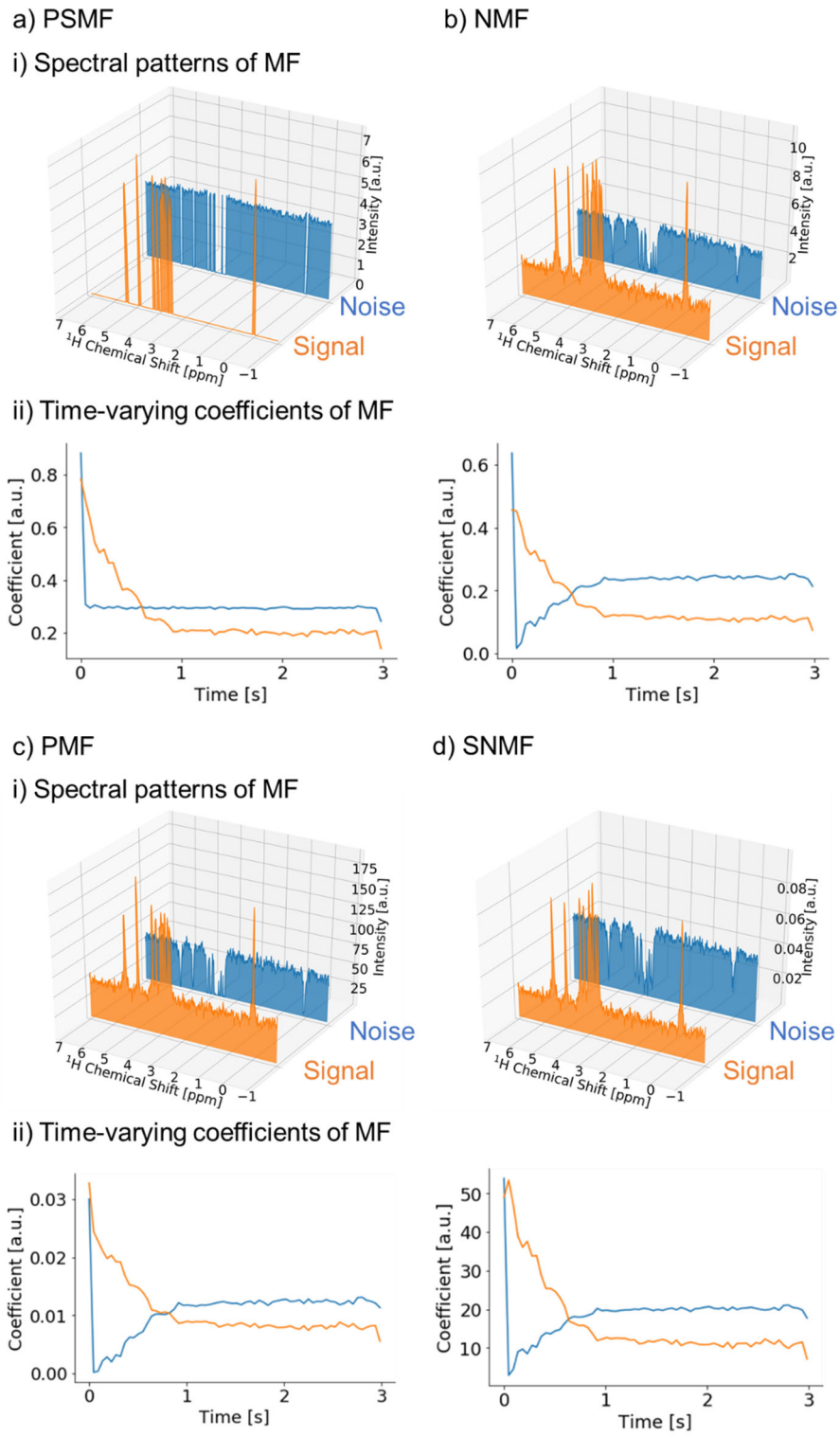


Figure S4. Comparison of four types of matrix factorization (MF) for signal deconvolution. MF was performed using four different methods, PSMF (a), NMF (b), PMF (c), and SNMF (d), and the effect on separated components was investigated. i) Spectral patterns of each MF method. ii) Time-varying coefficient of each component in each MF method.

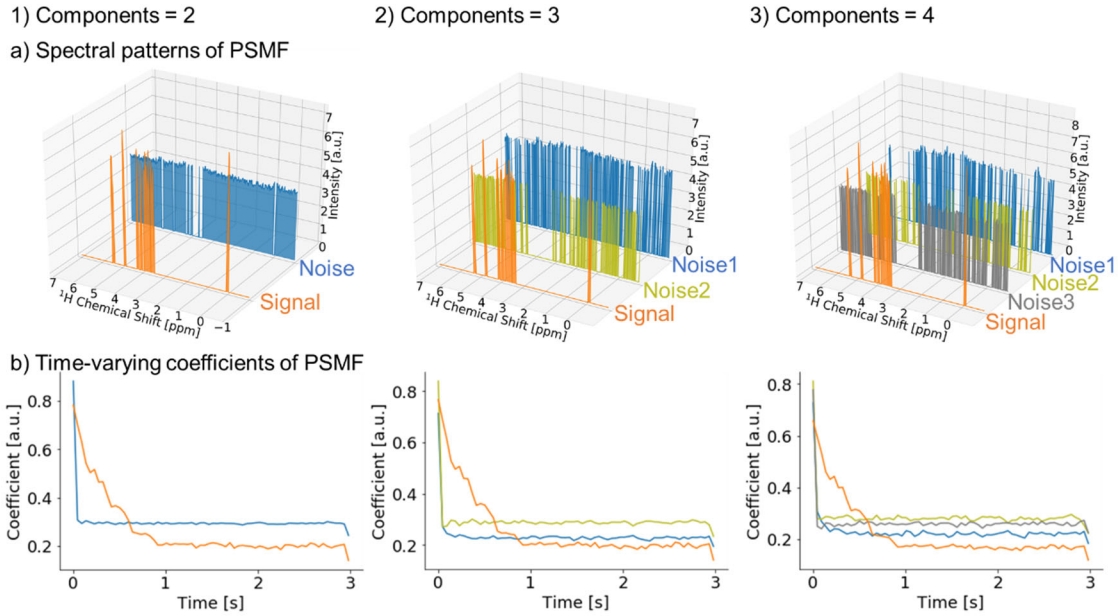


Figure S5. Effect of the number of components in PSMF. PSMF was performed using different numbers of components, two components (1), three components (2), and four components (3), and the effect on separated components was investigated. a) Spectral patterns of PSMF. b) Time-varying coefficient of each component in PSMF.

a) Heatmap of NMR data using CPMG

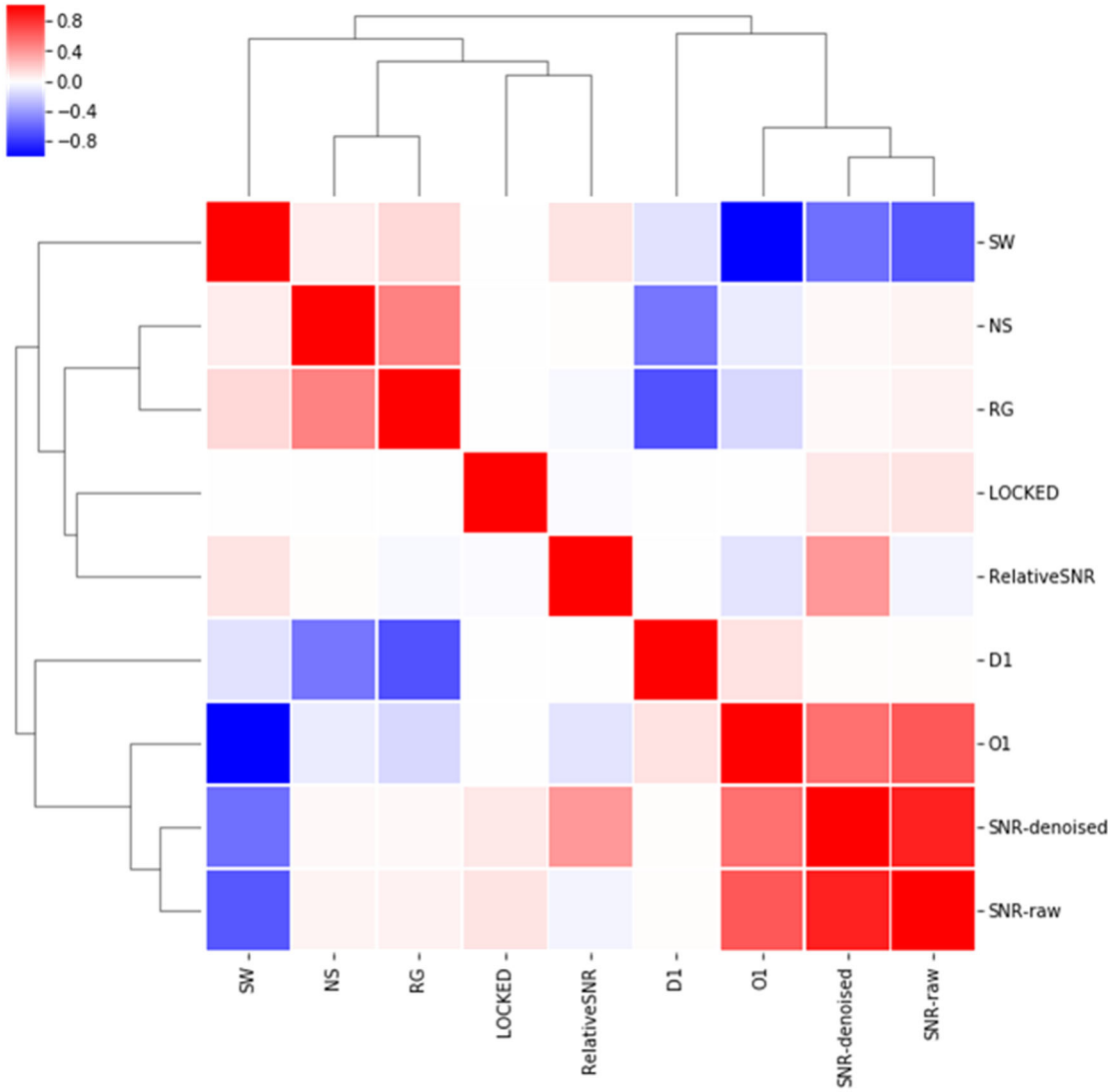


Figure S6a. Relationship between SNR and acquisition parameters of NMR data using CPMG. a) Heatmap. In the network diagram, positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is indicated by edge thickness. Abbreviations: SNR-raw, SNR of raw data; SNR-denoised, SNR of denoised data; RelativeSNR, relative SNR; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; O1, the offset of the transmitter frequency; LOCKED, if LOCK is on, value is 1, if not, value is 0.

b) Network diagram of NMR data using CPMG

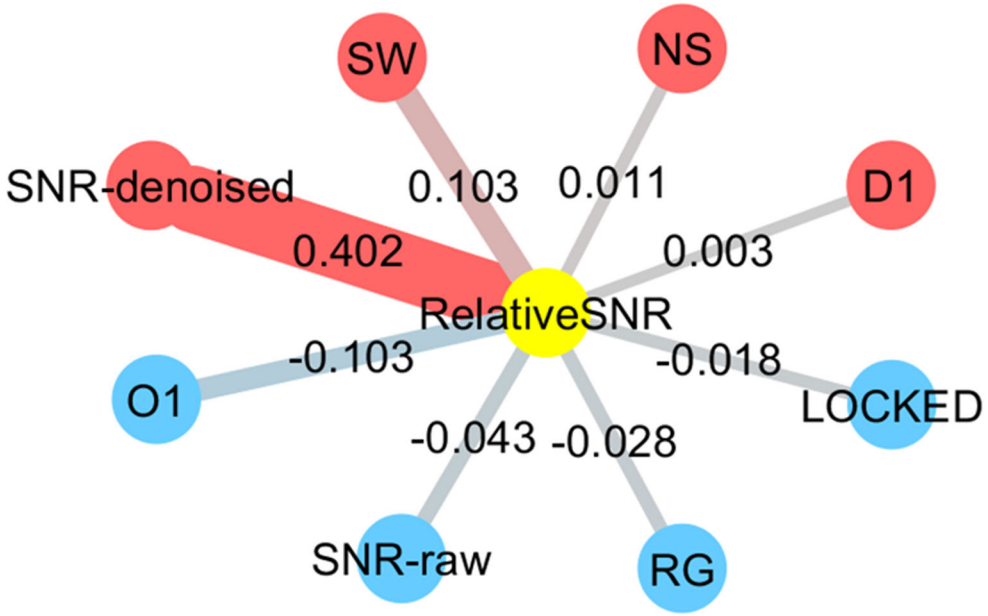


Figure S6b. Relationship between SNR and acquisition parameters of NMR data using CPMG. b) Network diagram. In the network diagram, positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is indicated by edge thickness. Abbreviations: SNR-raw, SNR of raw data; SNR-denoised, SNR of denoised data; RelativeSNR, relative SNR; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; O1, the offset of the transmitter frequency; LOCKED, if LOCK is on, value is 1, if not, value is 0.

a) Heatmap of NMR data using WATERGATE

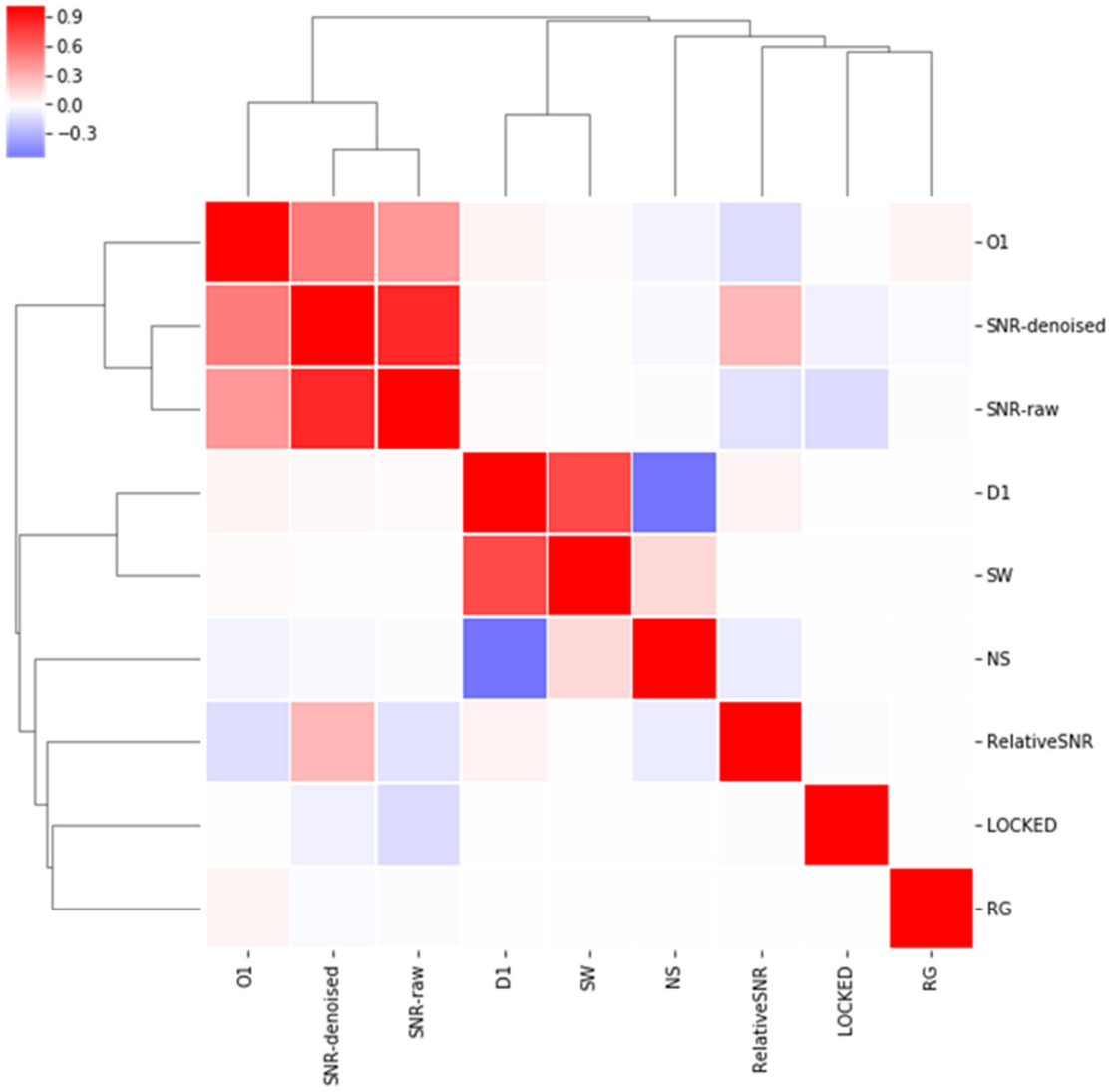


Figure S7a. Relationship between SNR and acquisition parameters of NMR data using WATERGATE. a) Heatmap. In the network diagram, positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is indicated by edge thickness. Abbreviations: SNR-raw, SNR of raw data; SNR-denoised, SNR of denoised data; RelativeSNR, relative SNR; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; O1, the offset of the transmitter frequency; LOCKED, if LOCK is on, value is 1, if not, value is 0.

b) Network diagram of NMR data using WATERGATE

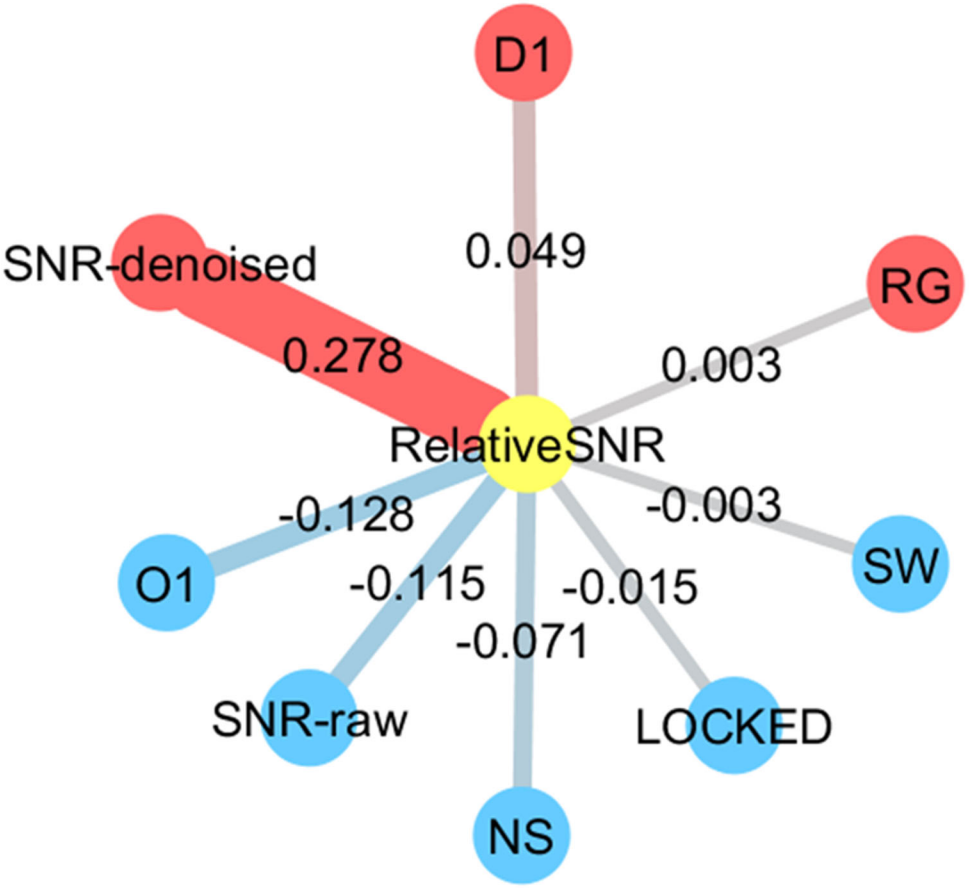


Figure S7b. Relationship between SNR and acquisition parameters of NMR data using WATERGATE. b) Network diagram. In the network diagram, positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is indicated by edge thickness. Abbreviations: SNR-raw, SNR of raw data; SNR-denoised, SNR of denoised data; RelativeSNR, relative SNR; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; O1, the offset of the transmitter frequency; LOCKED, if LOCK is on, value is 1, if not, value is 0.

a) Heatmap of diffusion-edited NMR

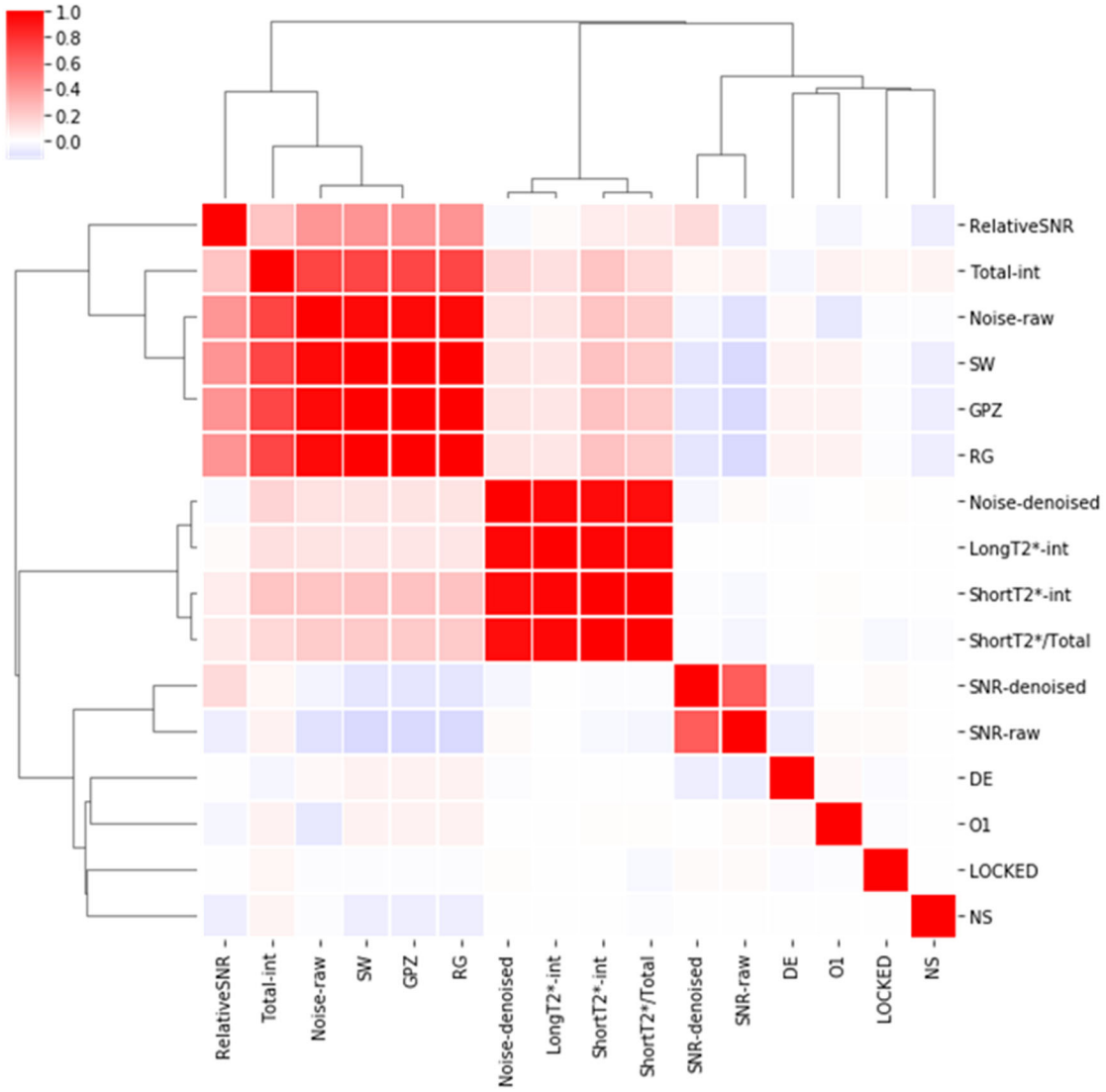


Figure S8a. Relationship between the data quality (SNR and the composition of the separated signal) and acquisition parameters of diffusion-edited NMR. a) Heatmap. In the network diagram, positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is indicated by edge thickness. Abbreviations: SNR-raw, SNR of raw data; SNR-denoised, SNR of denoised data; RelativeSNR, relative SNR; Total-int, total intensity; ShortT2*-int, intensity of short T_2^* signal; LongT2*-int, intensity of long T_2^* signal; ShortT2*/Total, ratio of intensity of long T_2^* signal to total intensity; Noise-raw, noise of raw data; Noise-denoised, noise of denoised data; GPZ, gradient pulse in the *z*-axis; RG, receiver gain; NS, number of scans; DE, pre-scan delay; SW, spectral width; O1, the offset of the transmitter frequency ; LOCKED, if LOCK is on, value is 1, if not, value is 0.

b) Network diagram of diffusion-edited NMR

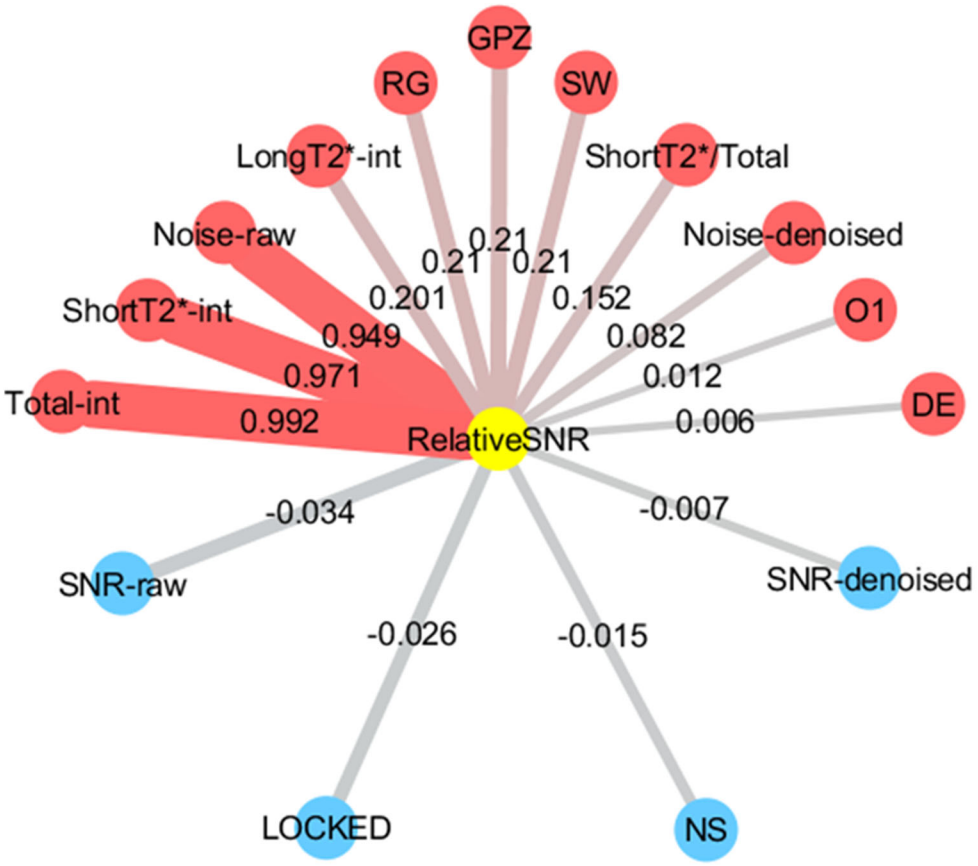


Figure S8b. Relationship between the data quality (SNR and the composition of the separated signal) and acquisition parameters of diffusion-edited NMR. b) Network diagram. In the network diagram, positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is indicated by edge thickness. Abbreviations: SNR-raw, SNR of raw data; SNR-denoised, SNR of denoised data; RelativeSNR, relative SNR; Total-int, total intensity; ShortT2*-int, intensity of short T2* signal; LongT2*-int, intensity of long T2* signal; ShortT2*/Total, ratio of intensity of long T2* signal to total intensity; Noise-raw, noise of raw data; Noise-denoised, noise of denoised data; GPZ, gradient pulse in the *z*-axis; RG, receiver gain; NS, number of scans; DE, pre-scan delay; SW, spectral width; O1, the offset of the transmitter frequency ; LOCKED, if LOCK is on, value is 1, if not, value is 0.

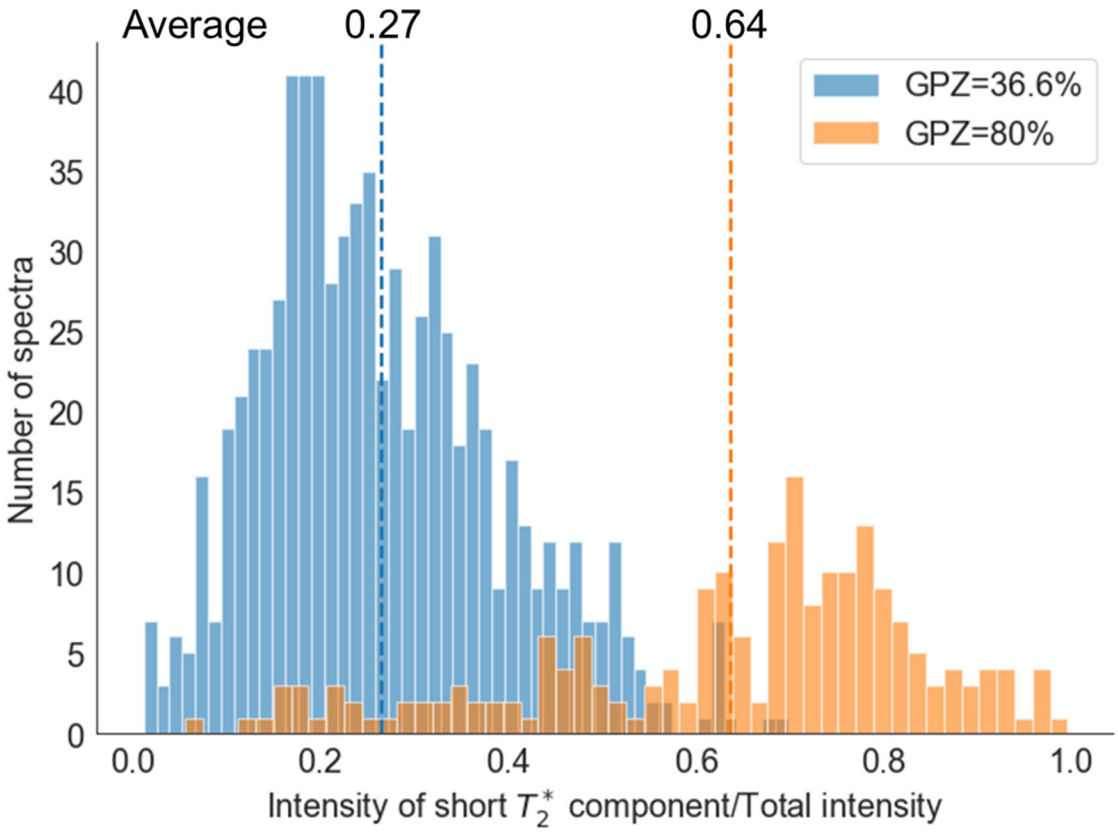


Figure S9. Histogram of the composition of the separated signal in diffusion-edited NMR data. We investigated the relationship between the composition of the separated signal and the gradient pulse in the *z*-axis (GPZ) parameter of diffusion-edited NMR. The histogram shows the relative SNR in NMR data measured using two different GPZ values. Shown is the ratio of the sum of short T_2 intensity to total intensity for GPZ = 36.6% (blue) and for GPZ = 80% (red). The average value in each pulse sequence is indicated.

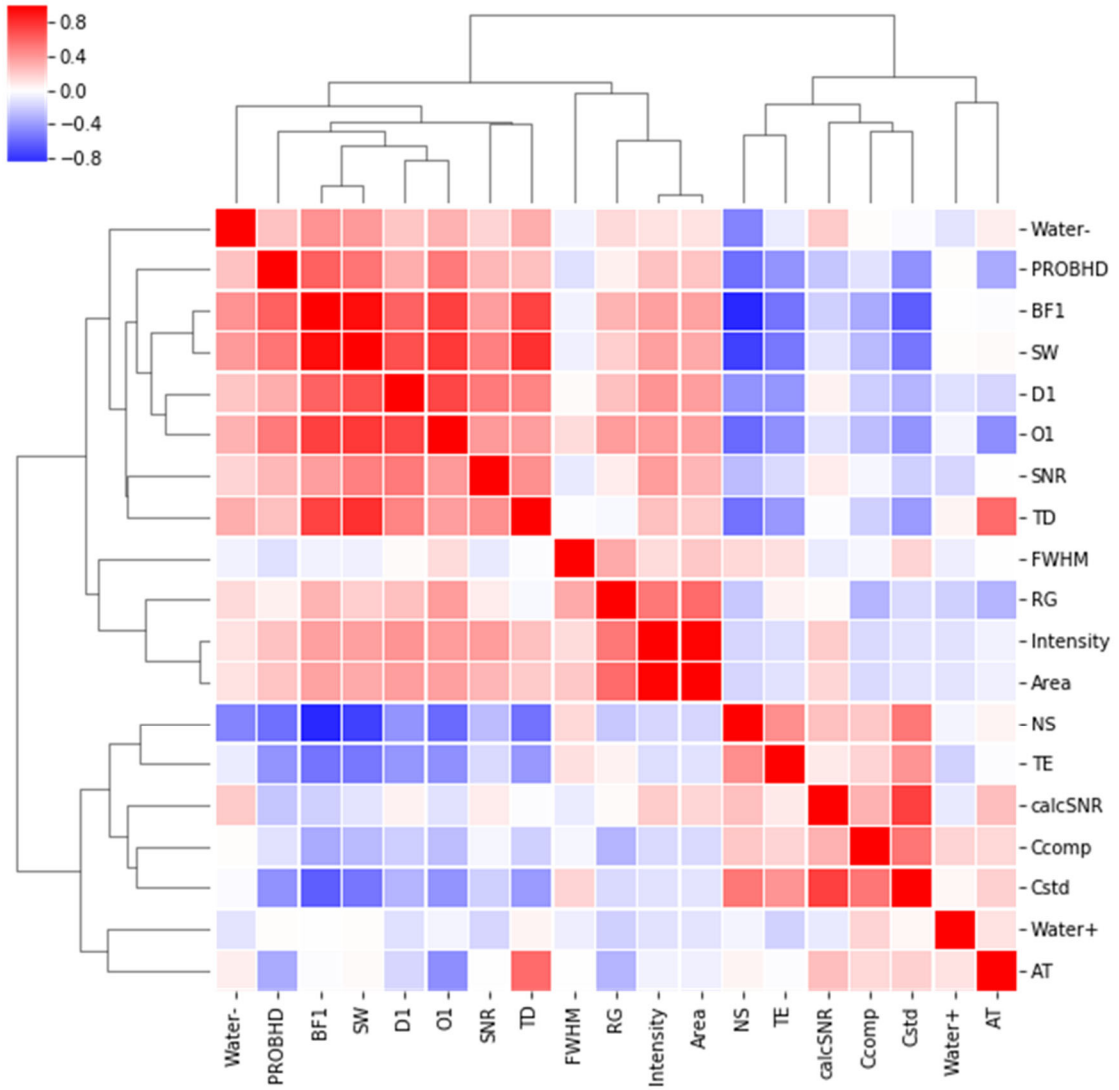


Figure S10. Heatmap summarizing correlation analysis between the data quality (SNR and signal values) and experimental parameters. Positive correlations are red; negative correlations are blue; and the magnitude of the correlation coefficient is shown as a color gradient. The parameters are clustered according to the similarity of their correlation coefficient as determined by hierarchical cluster analysis. Abbreviations: SNR, signal to noise ratio; calcSNR, calculated SNR; Cstd, concentration of standard compound; Ccomp, concentration of compound; Water+, positive intensity of water signal peak to standard peak; Water-, negative intensity of water signal peak to standard peak; Intensity, intensity of standard signal; FWHM, full width at half maximum; Area, area of standard signal; RG, receiver gain; NS, number of scans; D1, relaxation delay time; SW, spectral width; AT, acquisition time; TD, time-domain data size; O1, offset of transmitter frequency; TE, temperature; BF1, basic transmitter frequency for channel F1 in Hertz; PROBHD, if cryoprobe, value is 4, if not, value is 0.

Table S1. Original and denoised parameters and spectral values in citric acid data

^{1}H Chemical shift (ppm)		2.67	2.64	2.55	2.53	0
J value (Hz)		15.0		15.0		—
Original	Peak intensity	134991427	195407552	214161581	147849699	107410280
	FWHM (Hz)	2.40	2.48	2.31	2.26	2.13
Denoised	Peak intensity	134842313	194951941	213631581	147369942	107465227
	FWHM (Hz)	2.40	2.48	2.32	2.27	2.13
Error	Peak intensity (%)	0.11	0.23	0.25	0.32	-0.05
	FWHM (%)	0.02	-0.05	-0.52	-0.02	-0.04

^{1}H chemical shift, *J* value, peak intensity, and full width at half maximum (FWHM) are shown as the values of the original spectrum and the denoised spectrum in citric acid. Errors were calculated the difference between the original spectral value and the denoised spectral value. Relative SNR of this spectra is 1.14-fold.

Table S2. Summary of NMR spectra derived from sample ID of 1 to 10

Sample ID	PULPROG	D1	DE	NS	O1	RG	SW	TD	SNR-denoised	SNR-raw	Relative SNR	AT
1	CPMG	2	10	32	3457	108	14	32768	38229.09	14033.33	2.72	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	667.75	323.26	2.07	0.73
	Watergate	2.5	10	32	3295	108	14	32768	22718.07	5850.79	3.88	1.67
2	CPMG	2	10	32	3457	108	14	32768	1517567.61	504865.23	3.01	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	397.92	456.65	0.87	0.73
	Watergate	2.5	10	32	3295	108	14	32768	34829.85	8669.44	4.02	1.67
3	CPMG	2	10	32	3457	108	14	32768	1262994.59	443656.14	2.85	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	137.17	194.33	0.71	0.73
	Watergate	2.5	10	32	3295	108	14	32768	11642.91	4351.65	2.68	1.67
4	CPMG	2	10	32	3457	108	14	32768	102173.72	34671.49	2.95	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	679.61	246.47	2.76	0.73
	Watergate	2.5	10	32	3295	108	14	32768	15930.79	4331.21	3.68	1.67
5	CPMG	2	10	32	3457	108	14	32768	174450.86	77819.70	2.24	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	263.71	254.43	1.04	0.73
	Watergate	2.5	10	32	3295	108	14	32768	27185.68	6901.45	3.94	1.67
6	CPMG	2	10	32	3457	108	14	32768	155495.88	42460.54	3.66	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	617.51	306.23	2.02	0.73
	Watergate	2.5	10	32	3295	108	14	32768	15631.96	4608.96	3.39	1.67
7	CPMG	2	10	32	3457	108	14	32768	62782.12	29865.18	2.10	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	270.25	261.18	1.03	0.73
	Watergate	2.5	10	32	3295	108	14	32768	33748.08	7605.11	4.44	1.67

Sample ID	PULPROG	D1	DE	NS	O1	RG	SW	TD	SNR-denoised	SNR-raw	Relative SNR	AT
8	CPMG	2	10	32	3457	108	14	32768	100221.74	19528.38	5.13	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	1121.33	406.83	2.76	0.73
	Watergate	2.5	10	32	3295	108	14	32768	38506.44	7167.94	5.37	1.67
9	CPMG	2	10	32	3457	108	14	32768	54878.55	22587.59	2.43	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	1158.86	581.15	1.99	0.73
	Watergate	2.5	10	32	3295	108	14	32768	18295.45	7211.20	2.54	1.67
10	CPMG	2	10	32	3457	108	14	32768	58250.19	18693.05	3.12	1.67
	Diffusion-edited	2	10	128	3291	388	16	16384	271.59	265.08	1.02	0.73
	Watergate	2.5	10	32	3295	108	14	32768	32553.16	10245.24	3.18	1.67

Table S2 provides sample title, solvent and acquisition time, acquisition point, and original SNR as information about the sample and acquisition parameters. All data is available at http://dmar.riken.jp/NMRinformatics/SIforDCTN.zip. Abbreviations: PULPROG, pulse program used for the acquisition; D1, relaxation delay time; DE, pre-scan delay; NS, number of scans; O1, offset of transmitter frequency; RG, receiver gain; SW, spectral width; TD, time-domain data size; SNR-denoised, SNR of denoised data; SNR-raw, SNR of raw data; RelativeSNR, relative SNR; AT, acquisition time.

Table S3. FID datasets used for noise factor analysis

NMR	Benchtop NMR		High-field NMR								
	60 MHz		500 MHz				600 MHz			700 MHz	
Source	RIKEN	NUIS	RIKEN	BMRB	BML	HMDB	RIKEN	BMRB	HMDB	RIKEN	BMRB
Glucose	nanalysis (NMReady60PRO)	nanalysis (NMReady60PRO)	Bruker (c6-500c)	Bruker (MMC) [3]	Bruker (BML)	—	—	Bruker (MMC)	Varian (HMDB)	Bruker (c6-700b) [2]	Bruker (NIST)
Sucrose	nanalysis (NMReady60PRO)	nanalysis (NMReady60PRO)	Bruker (c6-500c)	Bruker (MMC) [2]	Bruker (BML) [2]	Varian (HMDB)	—	Bruker (MMC)	—	Bruker (c6-700b) [2]	Bruker (NIST)
Citric acid	nanalysis (NMReady60PRO)	nanalysis (NMReady60PRO)	Bruker (c6-500c)	Bruker (MMC) [3]	Bruker (BML)	Varian (HMDB)	—	—	—	Bruker (c6-700b) [2]	—
Lactic acid	nanalysis (NMReady60PRO)	nanalysis (NMReady60PRO)	Bruker (c6-500c)	Bruker (MMC) [3]	Bruker (BML)	Varian (HMDB) [2]	Bruker (c5-600c) [1]	Bruker (MMC)	—	Bruker (c6-700b)	Bruker (NIST)

We collected 48 sets of NMR data measured by low- and high-field NMR at multiple institutions to investigate the comprehensive relationship between noise and several acquisition parameters. Abbreviations: RIKEN, RIKEN Yokohama Campus; NUIS, Niigata University of International and Information Studies; BMRB, Biological Magnetic Resonance Data Bank; BML, Birmingham Metabolite Library; HMDB, Human Metabolome Database; MMC, Madison Metabolomics Consortium; NIST, National Institute of Standards and Technology. The NMR spectrometer manufacturer is listed; the product name, organization who generated the dataset, or control number is shown in parentheses. In the case of multiple data, the number of data used is indicated in square brackets.

3. References

1. Liu, H.; Dong, H.; Ge, J.; Liu, Z.; Yuan, Z.; Zhu, J.; Zhang, H. A fusion of principal component analysis and singular value decomposition based multivariate denoising algorithm for free induction decay transversal data. *Rev. Sci. Instrum.* **2019**, *90*, 035116
2. Keeler, J. Understanding NMR Spectroscopy. Appollo – University of Cambridge Repository: Cambridge, UK, 2004, doi:10.17863/CAM.1291.
3. Liu, H.; Dong, H.; Ge, J.; Bai, B.; Yuan, Z.; Zhao, Z. Research on a secondary tuning algorithm based on SVD & STFT for FID signal. *Meas. Sci. Technol.* **2016**, *27*, 105006.
4. Dueck, D.; Morris, Q.; Frey, B. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* **2005**, *21*, I144–I151.
5. Smith, J.O. *Mathematics of the Discrete Fourier Transform (DFT): with Audio Applications*, 2nd ed. W3K Publishing, **2007**.
6. Claridge, T. MNova: NMR data processing, analysis, and prediction software. *J Chem Inf Model* **2009**, *49*, 1136–1137.
7. Larive, C.K.; Jayawickrama, D.; Orfi, L. Quantitative analysis of peptides with NMR spectroscopy. *Appl. Spectrosc.* **1997**, 51, 1531–1536.

Appendix C

Supplementary Materials for

Signal Deconvolution and Generative Topographic Mapping Regression for Solid-state NMR of Multi-component Materials

Python tools developed in this study are available at http://dmar.riken.jp/NMRinformatics/.

Supplementary figures

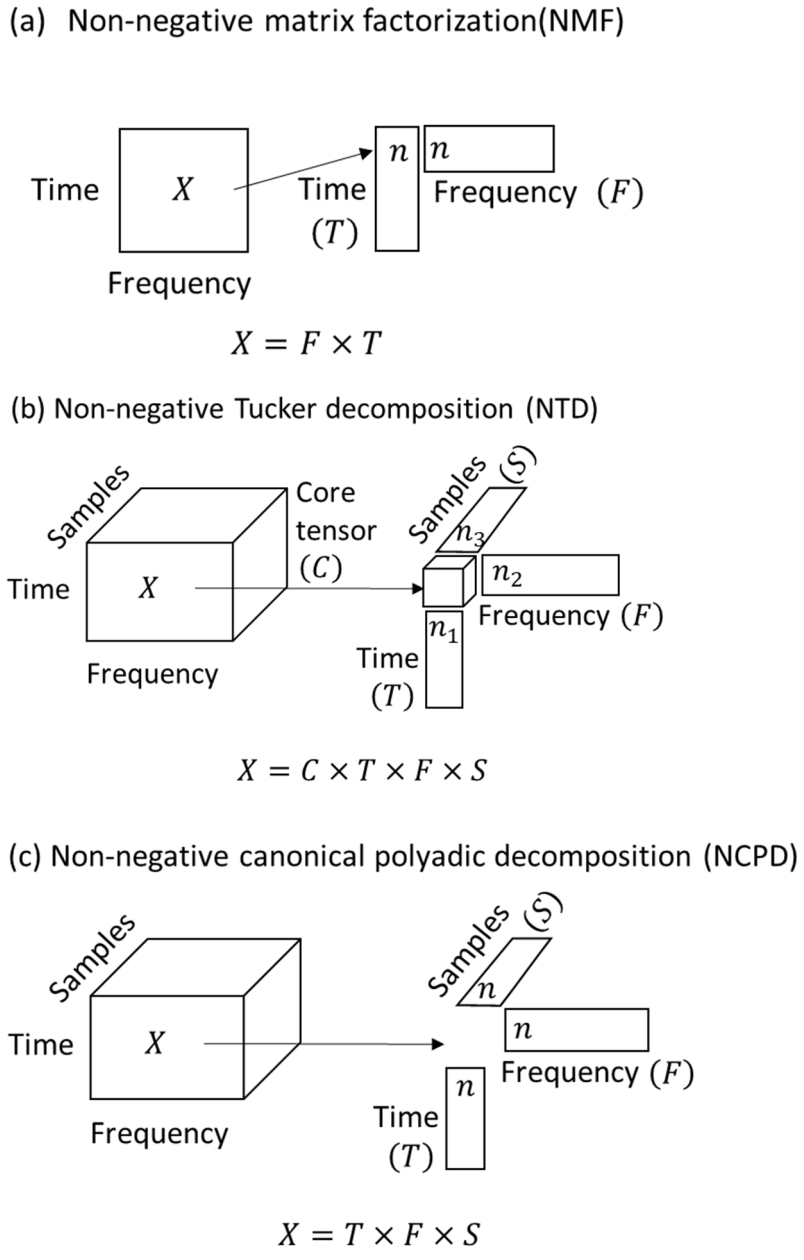
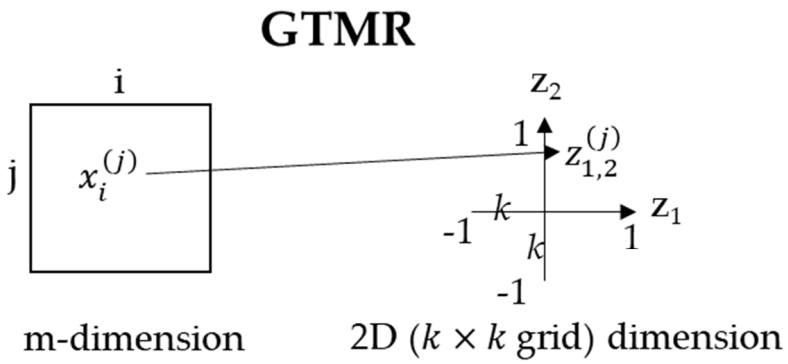


Figure S1. Algorithms of non-negative tensor/matrix factorization (NTF, NMF). (a) Non-negative matrix factorization (NMF). (b) Non-negative Tucker decomposition (NTD). (c) Non-negative canonical polyadic decomposition (NCPD). In the case of two-dimensional datasets such as a matrix with time and frequency axes, the FID is separated into each component based on factors of time and frequency by matrix factorization. For analysis of the three-dimensional dataset of multiple samples and parameters, tensor methods such as NTD and NCPD can be used.



$$x^{(i)} = f\left(z^{(j)}\right)$$

Figure S2. Algorithm of generative topographic mapping regression (GTMR). Using the GTMR, multi-dimensional and multi-component data can be mapped into the reduced dimensional space.

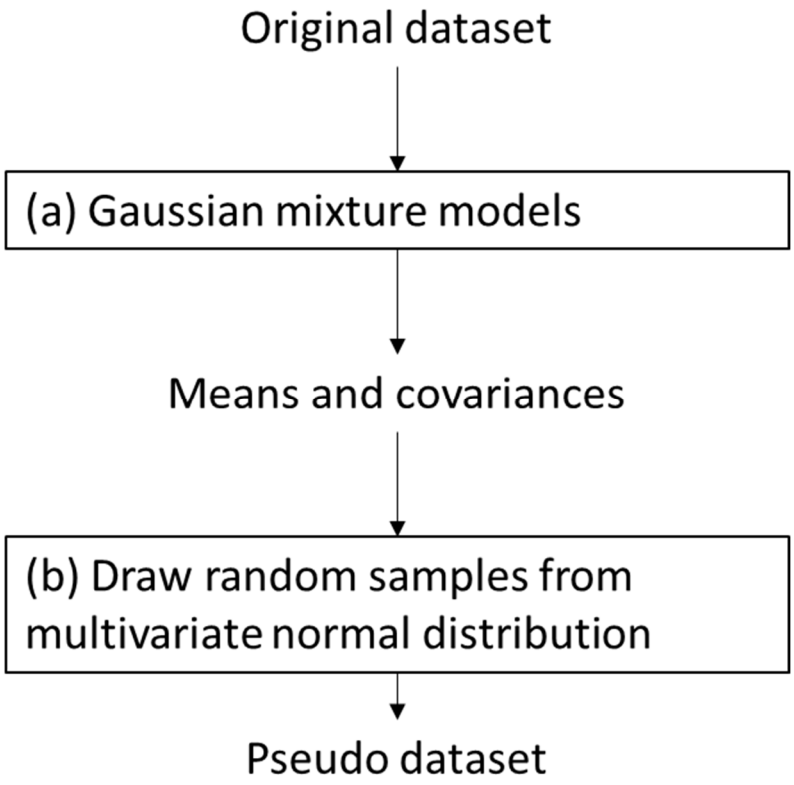


Figure S3. Algorithm of generating data using Gaussian mixture models (GMM). (a) GMM estimates the distribution of the dataset. (b) Draw random samples based on distribution estimated by GMM.

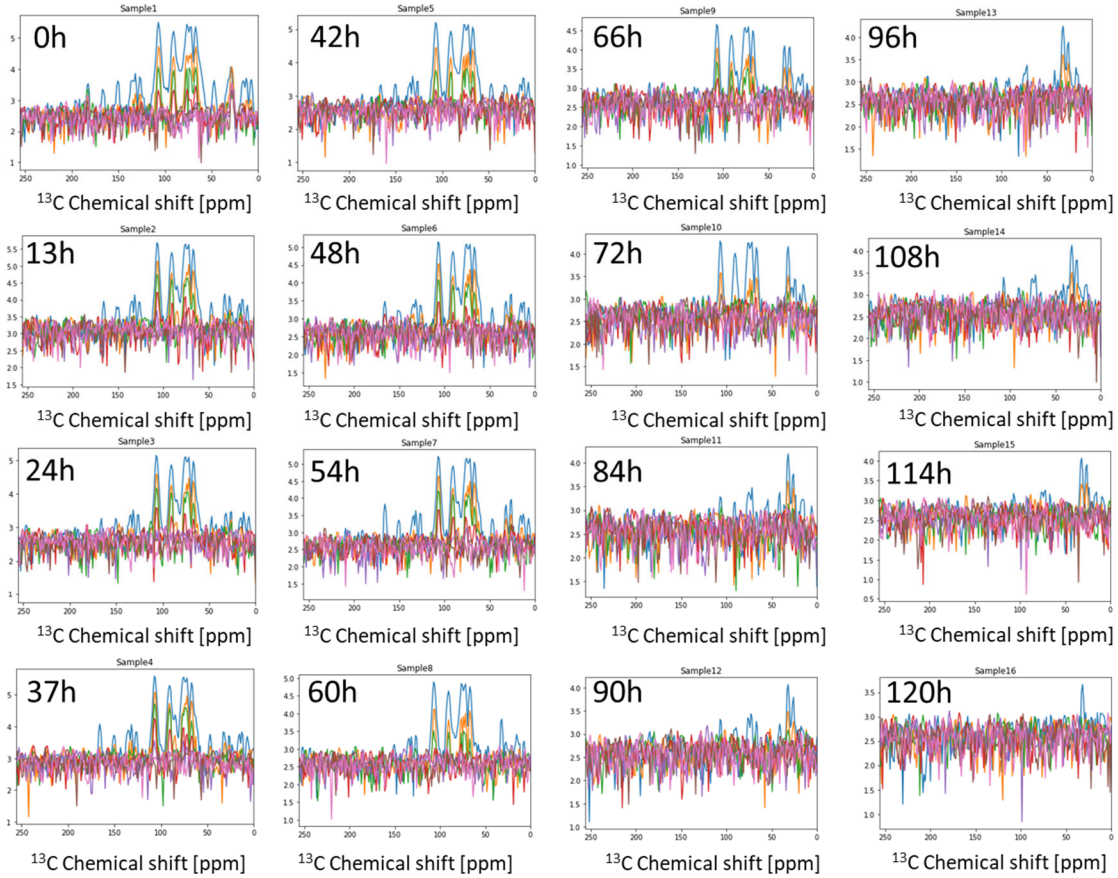


Figure S4. Short-time Fourier transform processed NMR (STFT-NMR) signals in ^{13}C CP-MAS of the cellulose degradation process. These figures show STFT processed NMR data for each time of the cellulose degradation process.

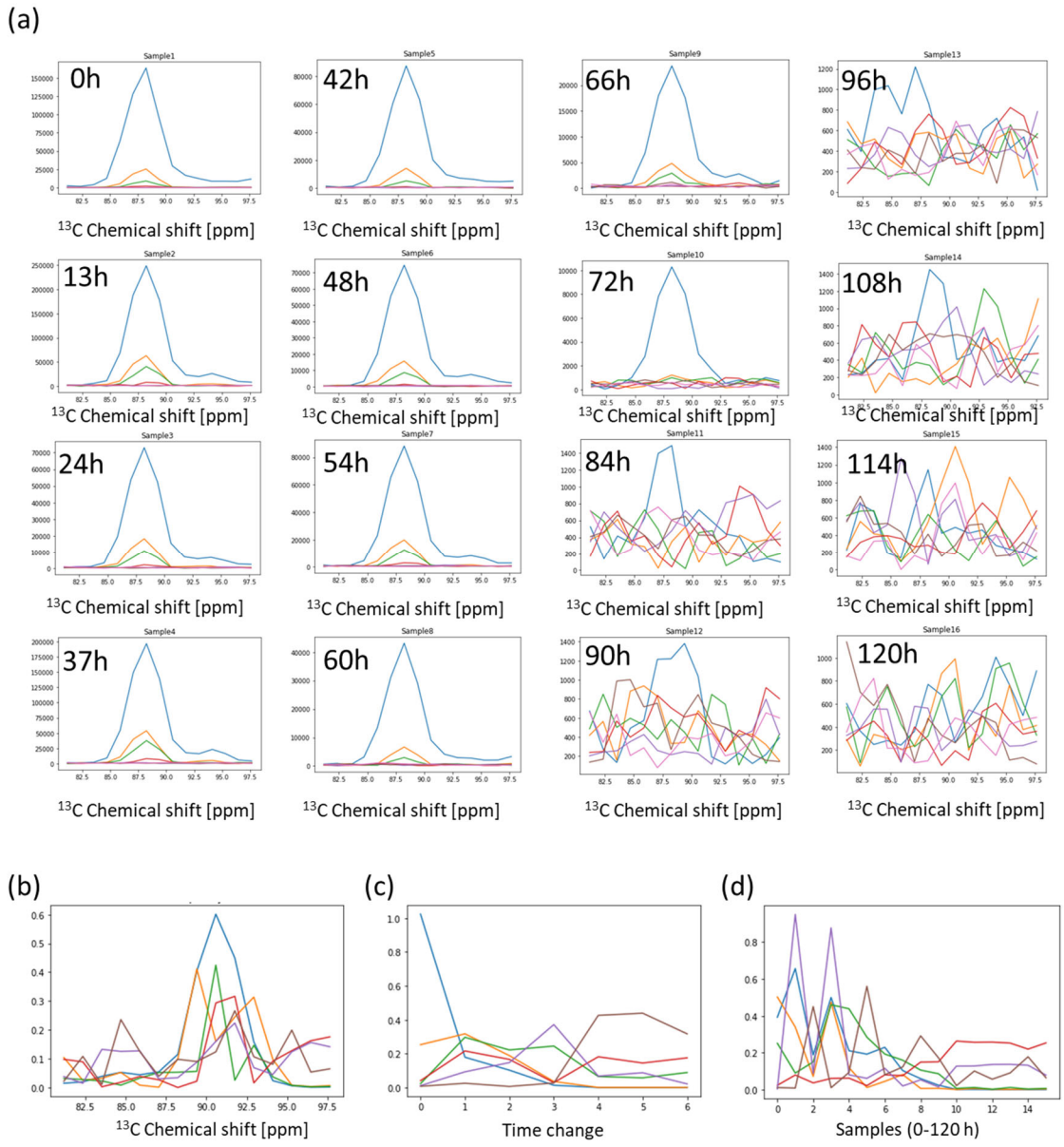


Figure S5. Signal deconvolution of cellulose C4 region using non-negative Tucker decomposition (NTD) in ^{13}C CP-MAS of cellulose degradation process. (a) These figures show cellulose C4 region STFT processed NMR data for each time of the cellulose degradation process. The figures (b-d) show spectral patterns (b), time change of separated components (5c), and composition of separated components (d) as results of separating the spectrum of cellulose C4 region into six components.

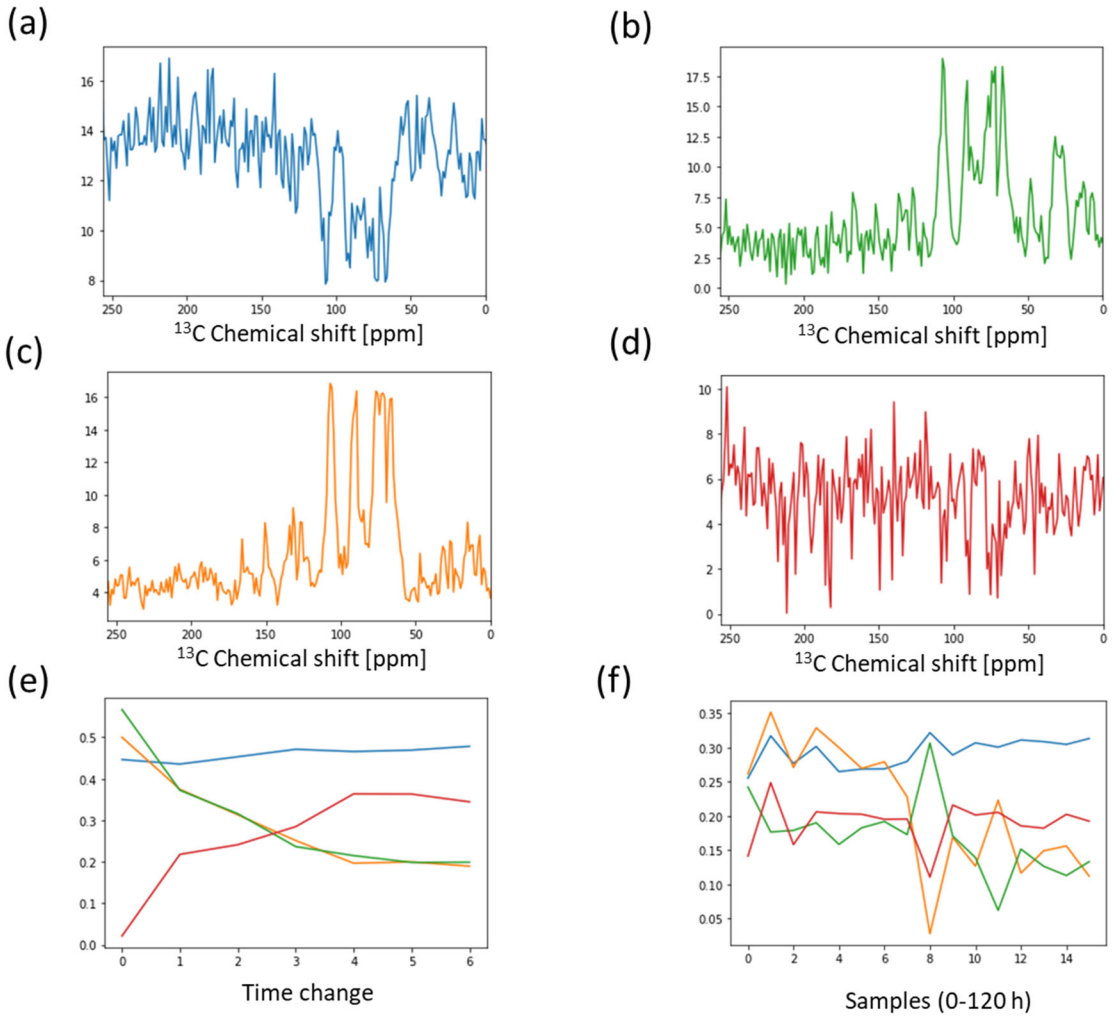


Figure S6. Signal deconvolution using non-negative canonical polyadic decomposition (NCPD) in ^{13}C CP-MAS of the cellulose degradation process. These figures show spectral patterns (a-d), time change of separated components (e), and composition of separated components (f) as results of separating the spectrum of cellulose using NCPD.

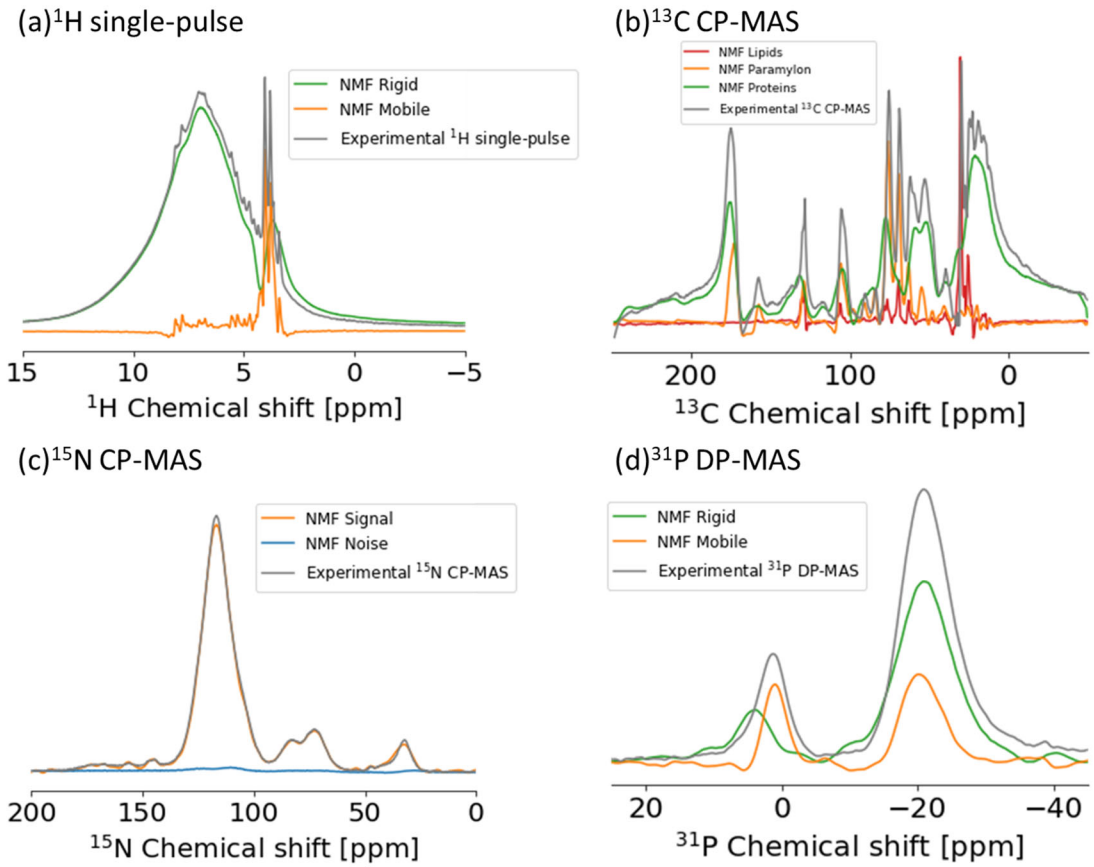


Figure S7. Signal deconvolution using MF to various NMR spectra in *E.graclis* samples. These figures show results of the signal deconvolution method using NMF for ^{1}H (a), ^{13}C (b), ^{15}N (c) and ^{31}P (d) spectra of microalgae such as *E. gracilis* in a multi-component system.

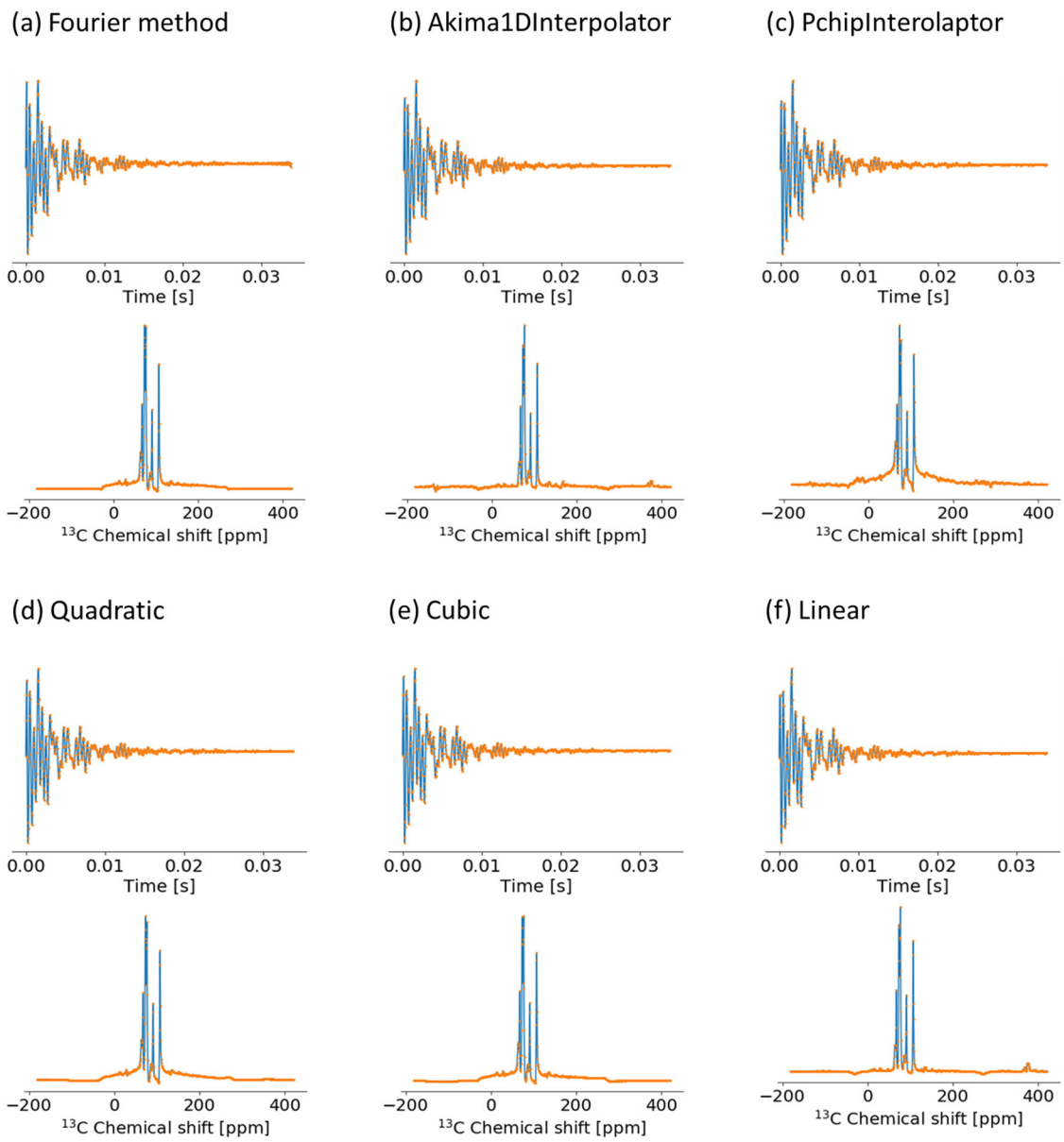


Figure S8. Application of interpolation methods for signal deconvolution of NMR data with insufficient data points. These figures show results of the resampling method using Fourier method (a) and other interpolation methods such as Akima, PCHIP (Piecewise Cubic Hermite Interpolating Polynomial), quadratic, cubic and linear (b-f).

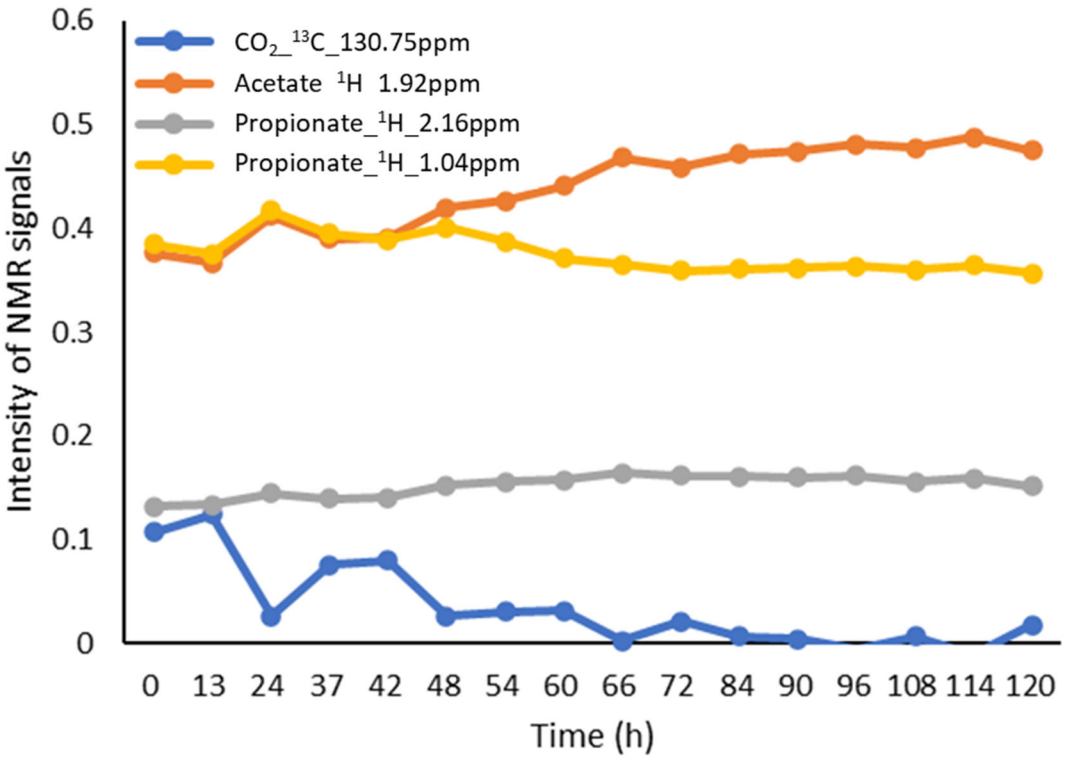


Figure S9. Summary of NMR signals for prediction in the cellulose degradation process. This figure shows the cellulose degradation process such as CO_2 (^{13}C chemical shift is 130.75 ppm), acetate (^{1}H chemical shift is 1.92 ppm), propionate (^{1}H chemical shift is 2.16 and 1.04 ppm) was captured by solution NMR.

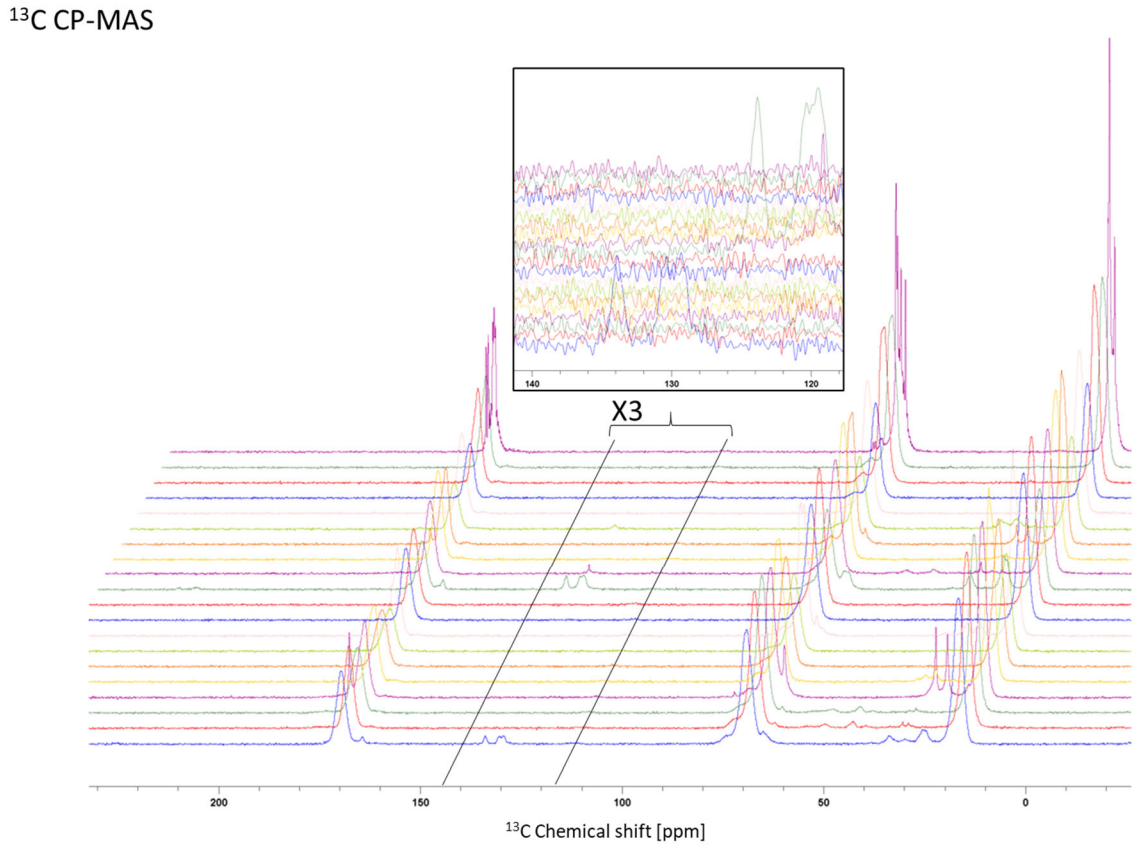


Figure S10. Summary of NMR data for prediction in polylactic acid (PLA). This figure shows ^{13}C CP-MAS spectra of 22 plastics.

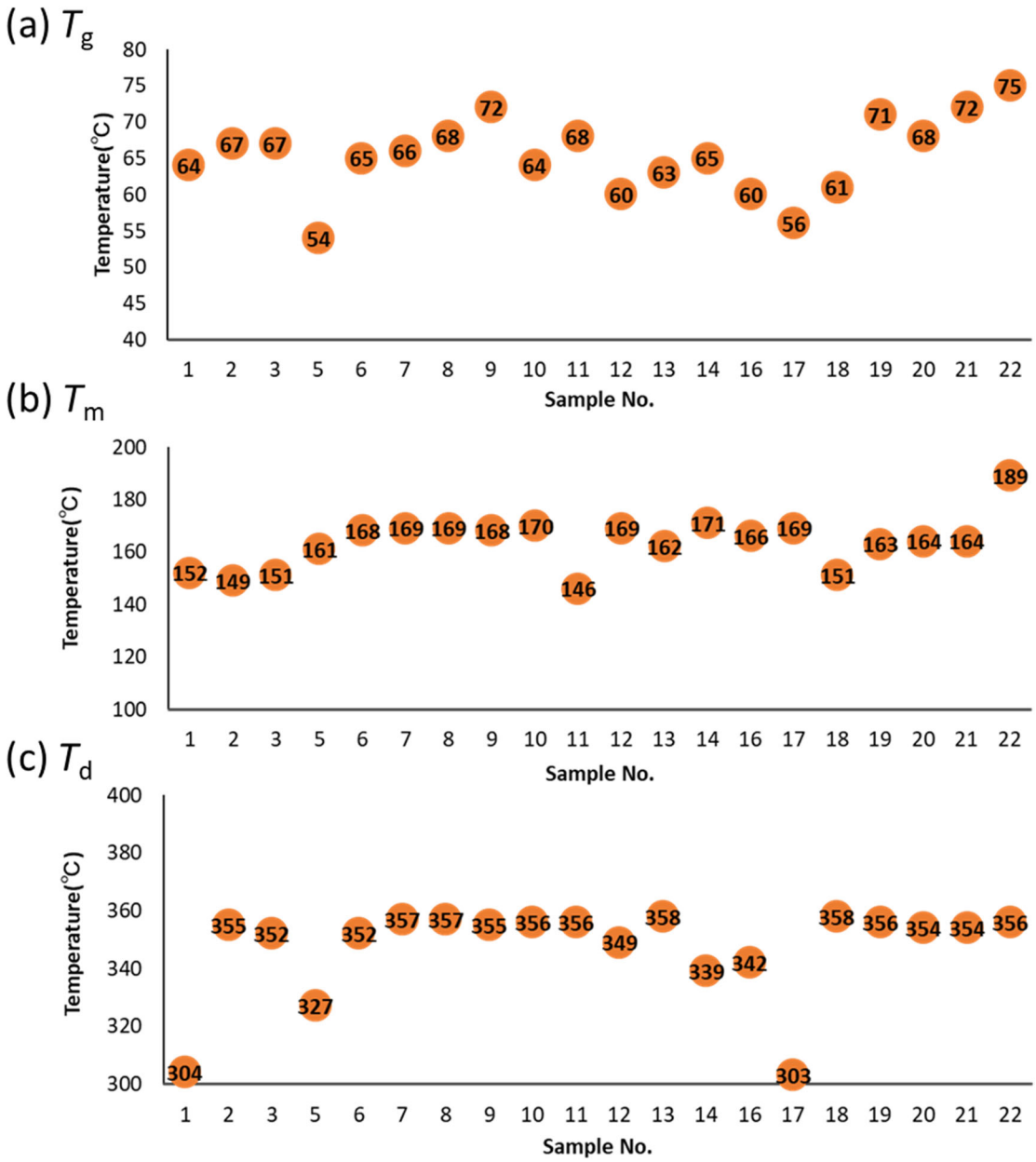


Figure S11. Summary of thermal analysis data for prediction in PLA. These figures show thermal analysis data of T_g (a), T_m (b), T_d (c) in 22 plastics.

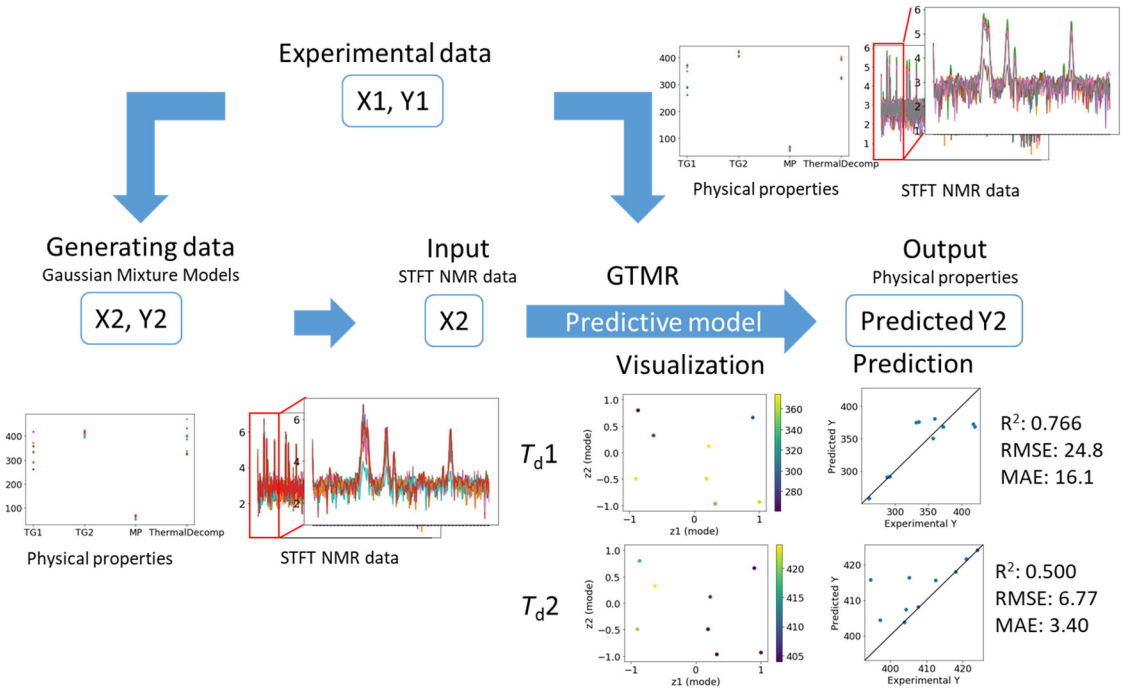


Figure S12. Prediction to thermal properties from NMR signals generated Gaussian mixture models (GMM) in poly-ε-caprolactone. This figure shows a scheme and result of predicting the thermal properties such as the degradation temperature (T_d)from pseudo ^{13}C CP-MAS spectra using GMM.