Electronic Supporting Information

Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples

Francesc Puig-Castellví,[a] Yolanda Pérez,[b] Benjamín Piña,[a] Romà Tauler,[a] and Ignacio Alfonso[c]

[a.] *Prof. R. Tauler, F. Puig-Castellví, Dr. B. Piña, Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain.*
[b.] *Dr. Y. Pérez, NMR Facility, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain.*
[c.] *Dr. I. Alfonso, Department of Biological Chemistry and Molecular Modelling, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain. E-mail: ignacio.alfonso@iqac.csic.es*

Table of Contents

Experimental Procedures

1.Preparation of NMR samples of synthetic mixtures

NMR samples were prepared in 600 µL of 0.25 mM sodium phosphate deuterated buffer (pH 7.0) with 200 mM DSS. Mixtures were prepared following the concentrations indicated in the table below:

Table S1. Metabolite concentrations used.

Sample name	DSS (in mM)	D-glucose (in mM)	Uridine (in mM)	L-glutamine (in mM)
Gln	0.25	0	0	1.6
2Gln/Glc	0.25	0.35	0	3.2
Uri	0.25	0	0.24	0
Gln/Uri	0.25	0	0.24	1.6
Gln/2Uri	0.25	0	0.48	1.6
Glc	0.25	0.35	0	0
Gln/Glc	0.25	0.35	0	1.6
Glc/Uri	0.25	0.35	0.24	0
Gln/Glc/Uri	0.25	0.35	0.24	1.6
2Glc/Uri	0.25	0.7	0.24	0

2.Preparation of NMR samples of yeast extracts

Yeast Growth. *S. cerevisiae S288C* cells were pre-cultured in YPD (1 % yeast extract, 1 % peptone, 2 % glucose) medium on an orbital shaker (150 rpm) at 30 ˚C overnight. All following cultures were cultured with these shaking and temperature conditions. 2 L of YNB Synthetic Complete medium (YSC, 1.7 g/L Yeast Nitrogen Base without amino acids and sulfate (Difco), 5 g/L $(NH_4)_2SO_4$) were inoculated with 200 µL of the pre-culture sample and left at the same temperature and shaking conditions until the culture reached an absorbance at 600 nm (A_{600}) of approximately 0.8 - 1. Pellets from these resulting cultures were collected by centrifuging the cultures, but not washed, at 2000 rpm for 3 min and 4 ˚C. Pellets were used right after for inoculating Erlenmeyer's containing either YSC medium or YPD medium.

Sample collection. 100 ml aliquots of every culture were collected six times during three days (0h, 2h, 4h, 6h, 10h, 24h, 48h and 72h). Samples were arrested with a cold shock in ice and cell were harvested by centrifugation at 4000 g for 3 min, discarding the supernatant. Cells were washed twice in 100 mM Na_2HPO_4 pH 7.0 followed by a centrifugation at 4700 g for 3 min. Resulting pellets were stored at -80 ˚C and lyophilized.

Metabolite extraction. Metabolites were extracted by following the protocol published in a previous work. 1800 µL of a solution of methanol-chloroform 1:2 (4 ºC) were added to the pellet, followed by a vigorous vortexing. A cold shock was then applied to the pellets for 5 times using the following procedure: the pellets were submerged in liquid nitrogen for 1 minute and consequently thawing in ice for 2 minutes. 400 µl of water were added to create the biphasic system. After homogenization by vortexing, a 3 min centrifugation at 16,500 rpm and 4 ºC was carried out. The aqueous phase (upper part) was collected. This process was repeated and samples were freeze-dried afterwards.

NMR sample preparation. Aqueous metabolites samples were dissolved in 650 µL of deuterated phosphate buffer (25 mM Na_2DPO_4, pH 7.0) in D_2O with 0.2 mM DSS as internal standard. Samples were centrifuged at 9,168 g for 5 min and the supernatant was collected and introduced into 5 mm NMR tubes.

3.Preparation of NMR samples of complex molecules

Cyclosporin A (Sigma-Aldrich) sample was prepared by dissolving 12mg in 0.55 mL deuterated benzene. 0.4 mM ^{15}N labeled Ubiquitin sample (Sigma-Aldrich) was prepared by dissolving the protein in 50 mM phosphate buffer (pH 6.2, 90% H_20/D_2O). Peptide AcCNPHFDLEC (Genscript HK Limited) was dissolved in DMSO-d_6 (1.5 mg, 2.5 mM). Lyophilized recombinant double labeled ^{13}C-^{15}N natively unfolded protein fragment was dissolved in 400 µL of 20 mM acetate buffer with 50 mM NaCl (pH 5, 90% H_20/D_2O) and introduced in a 5 mm Shigemi NMR tube.

4.NMR adcquisition parameters

All NMR spectra were acquired at 298 K on a 500 MHz AvanceIII HD NMR spectrometer equipped with a TCI cryoprobe from Bruker. All pulse sequences used are from Bruker TopSpn3.5pl6.

Synthetic mixture samples set. 1D NOESY spectra were recorded using the *noesygppr1d* pulse sequence and the following parameters: 256 scans, 4 seconds of relaxation delay, spectral width of 10 kHz and an acquired spectral size of 32k (final spectral size of 64k). The 90º pulse width (between 8.5 and 10.5 µs) and presaturation power for water suppression were measured for every sample before experiment acquisition. ^{1}H-^{13}C HSQC NMR spectra were recorded using the *hsqcetgpprsisp2.2.be* pulse sequence and the following parameters: 12 scans, 3 seconds of relaxation delay, spectral width of 12.7 kHz and an acquired spectral size of 548 data points in f1 dimension and of 1,536 data points in f2 dimension. The spectra were phase and baseline corrected and referenced to the DSS reference peak. After zero-filling, final spectral size in the ^{1}H-^{13}C HSQC NMR spectra were 1,024 data points in f1 (^{13}C) dimension and 2,048 data points in f2 (^{1}H) dimension.

Yeast extracts samples set. 1D NOESY spectra were recorded using the *noesygppr1d* pulse sequence and the following parameters: 256 scans, 4 seconds of relaxation delay, spectral width of 10 kHz and an acquired spectral size of 32k (final spectral size of 64k). The 90º pulse width (between 8.5 and 10.5 µs) and presaturation power for water suppression were measured for every sample before experiment acquisition. 2D ^{1}H-^{13}C HSQC NMR spectra were recorded using the *hsqcetgpprsisp2.2.be* pulse sequence and the following parameters: 12 scans, 3 seconds of relaxation delay, spectral width of 20.7 kHz in f1 and 7.9 kHz in f2, and an acquired spectral size of 548 data points in f1 dimension and of 1,536 data points in f2 dimension. The spectra were phase and baseline corrected and referenced to the DSS reference peak. After zero-filling, final spectral size in the ^{1}H-^{13}C HSQC NMR spectra were 1,024 data points in f1 (^{13}C) dimension and 2,048 data points in f2 (^{1}H) dimension.

Cyclosporin A sample. 2D ^{1}H-^{1}H TOCSY NMR spectrum was recorded using the *mlevphpp* pulse sequence and the following parameters: 8 scans, 1 seconds of relaxation delay, pulse width of 7.5 µs, spectral width of 6 kHz in both dimensions, and an acquired spectral size of 128 data points in f1 (^{1}H) dimension and of 1,024 data points in f2 (^{1}H) dimension, resulting in a final spectral size of 1,024 data points per dimension.

Ubiquitin sample. 2D ^{1}H-^{15}N HSQC NMR spectrum was recorded using the *hsqcetf3gpsi* pulse sequence and the following parameters: 2 scans, 1 second of relaxation delay, pulse width of 8 µs, spectral width of 8 kHz for 1H channel and 1.7 kHz for 15N channel, and an acquired spectral size of 64 data points in f1 dimension and of 1,024 data points in f2 dimension. After zero-filling, final spectral size in the ^{1}H-^{15}N HSQC NMR spectra were 256 data points in f1 dimension and 2,048 data points in f2 dimension.

Peptide sample 2D ^{1}H-^{1}H TOCSY NMR and 2D ^{1}H-^{1}H ROESY NMR spectra were recorded using the *mlevtgp* and *roesyphpp.2* pulse sequences, respectively. For both pulse sequences, the following parameters were used: 8 scans, 2 seconds of relaxation delay, pulse width of 8.0 s, spectral width of 6.5 kHz in both dimensions, and an acquired spectral size of 256 data points in f1 dimension and of 1,024 data points in f2 dimension, resulting in a final spectral size of 1,024 data points per dimension.

Natively unfolded protein sample: 3D HNCO spectrum was acquired using *hncogp3d* pulse sequence and the following acquisition parameters: 8 scans, 1 second of relaxation delay, pulse width of 8 µs, spectral width of 8.1/1.8/2,0 kHz for 1H/15N/13C channels, and acquired spectral size of 2048/72/128 for 1H/15N/13C dimensions.

5. VOI filtering function

The VOI filtering function (*voi2D.m*), and their equivalent for phase-sensitive 2D NMR spectra (*voi2Df.m*) and for 3D NMR spectra (*voi3D.m*) were implemented in Matlab programming language, and they can be downloaded from https://github.com/f-puig/VOI.

voi2D (and *voi2Df*) can filter a ^{1}H-^{13}C HSQC NMR matrix of 1,024 x 2,048 data-points in less than 0.5 seconds (time measured in an Intel workstation with 2.40 GHz, 128 GB RAM and 6 cores). *voi3D* can filter a 3D HNCO NMR spectrum of 1024 x 256 x 256 data-points in less than 150 seconds (time measured in the same Intel workstation).

Four different outputs are generated from the application of *voi2D* on a 2D NMR spectrum:

1. *VOImatrix*: 3-row data matrix containing the vector of filtered intensities and the two vectors containing the two measured δ.
2. *filtered_NMR*: 2D NMR filtered data matrix of equal in size than the input 2D NMR matrix, but with zero values on those positions considered to be noise.
3. *indexes*: list of positions in the input 2D NMR matrix that contain the filtered intensities.
4. *peak_arrays*: lists of filtered positions for every cluster.

VOImatrix has the compressed 2D NMR matrix. *filtered_NMR* has the reconstructed 2D NMR spectrum and it is very convenient for representing contour plots. The *indexes* vector is used to create the matrix of 2 or more VOI-processed spectra (see section 6 below). Finally, *peak_arrays* is used to determine the number of clusters per spectrum and the number of points per cluster.

6. Combining 2 or more VOI-processed spectra.

To combine two or more VOI-processed spectra, we need first to ensure that the spectra have the same dimensions and that the ppm1 and ppm2 measured values are the same. Otherwise, the spectra will need to be interpolated first using one of the spectra of the dataset as a reference.

VOI algorithm is first applied separately for every 2D NMR spectrum. Next, all the lists of VOI positions (*indexes*) are combined into one long matrix, excluding all repeated instances, with the list of common and uncommon VOIs.

To generate the matrix of VOI-processed spectra, an empty matrix with the same number of rows as samples and with the same number of columns as VOIs is created. Then, the first row is filled with the vector of selected intensities from the first spectra, and this process is repeated for the remaining spectra. Finally, all negative values are converted to 0.

7.Setting-up the *threshold* and *minvoi* parameters.

To define the *threshold* level, the most practical option is by checking the signal intensities of the 2D NMR spectrum using only one dimension (either ppm1 or ppm2). This means that, if the intensity values are represented on the first dimension (ppm1), we will have as many plotted lines as measured ppm2 values. Examples of the used 2D NMR datasets are given below. From this representation, it is easy to stablish an intensity *threshold* value higher than the observed noise. The selection of the threshold for two 2D NMR spectra is shown in detail in **Fig1A** and **Fig1B**.

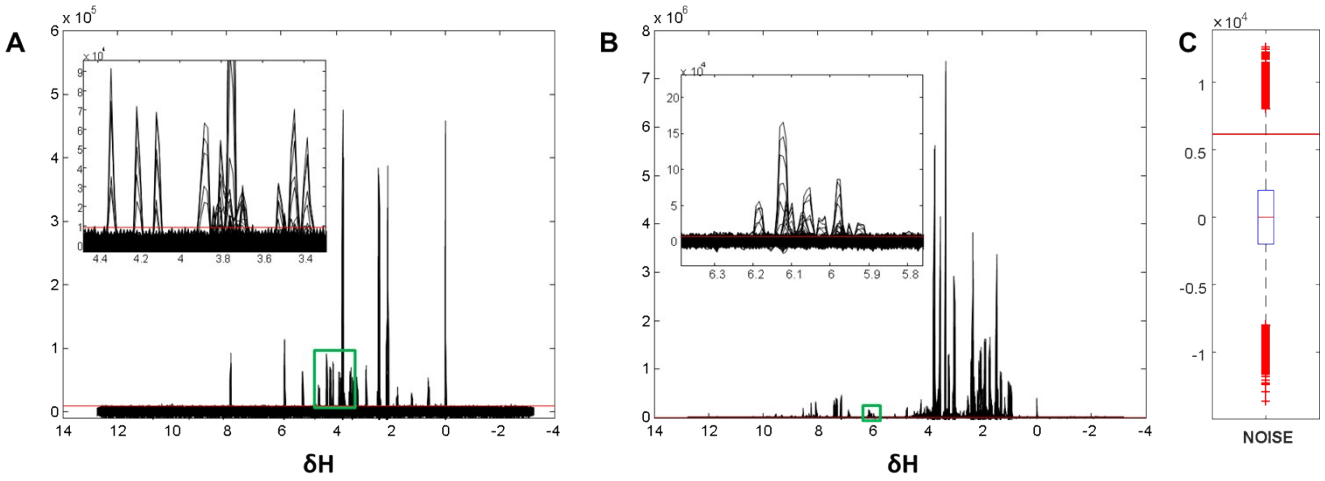

Figure S1. 2D NMR spectra plotted using only ppm1 (δ_H) dimension. A) Representative spectrum from dataset 1 (synthetic mixture). B) Representative spectrum from dataset 2 (yeast extract). C) Noise intensity values in dataset 2. Red horizontal line shows the applied threshold.

This threshold value can be also proposed from NMR regions where no meaningful resonances (only noise) are present.

By fixing the threshold level to the maximum intensity value detected in these NMR regions where only noise is present, filtered variables would only be representative of peak resonances. To avoid losing signal information related to the base of the signal resonances close to noise, the threshold value has to be decreased. For instance, in **Fig1C**, the maximum noise intensity measured was around 12,000, and the chosen threshold was decreased to 6,000. Most of the noise values comprised between 6,000 and 12,000 were also filtered after application of the *minvoi* (minimum number of adjacent points that define a peak) parameter. For all datasets tested in this work, fixing the threshold level to the half of the maximum noise value gave satisfactory filtering results.

For phase-sensitive spectra, two thresholds were used: one for peaks in phase (positive peaks) and another one for peaks in antiphase (negative peaks). Since noise values did not depend of the phase (noise is always centered to zero, see **Fig S1C**), the threshold level for both phases, in absolute values, was the same. Therefore, the threshold level was estimated in the positive phase and changed the

sign for the negative antiphase. To estimate the threshold level in the positive phase, the same procedure as for a phase-insensitive NMR spectrum described above was used.
To define the minimum number of adjacent points that define a NMR peak (*minvoi)*, 2D NMR spectra were investigated using the typical NMR MestreNova (Mestrelab Research S.L.) or Topspin (Bruker BioSpin GmbH) platforms. First, the smallest true peak was selected. In these two popular NMR platforms, such operation is rather fast, efficient and easy to use. Afterwards, under the MATLAB environment, the selected smallest peak was visualized, and the number of variables (or pixels) that define this peak are counted. The obtained value is the maximum recommended value for the *minvoi* parameter, which gives a satisfactory NMR signal filtering by adjusting this *minvoi* parameter by a a factor between 0.7 and 1.

8. NMR preprocessing

NMR spectra have been automatically referenced, phased and baseline corrected using TopSpin (Bruker, Germany) routines.
NMR preprocessing of ^{1}H NMR datasets. ^{1}H NMR Bruker files were imported to MestreNova v.11.0 (MestreLab Research), and an exponential apodization of 0.2 Hz was applied on each one of them. In MestreNova v.11.0 environment, spectra were converted into ASCII format and imported to Matlab R2016a (The Mathworks Inc. Natick, MA, USA). In Matlab, data was first normalized using Probabilistic Quotient Normalization (PQN) [1] using an in-house function, followed by a mean-centering using the PLS toolbox 8.2.0 (Eigenvecctor Research Inc., Wenatchee, WA, USA). Regions of water (4.41 - 5.16 ppm), methanol (3.30 - 3.37 ppm), chloroform (7.64 -7.69 ppm) and DSS (< 0.7 ppm) were removed. Data points which chemical shifts were higher than 9.7 ppm were also removed.
NMR preprocessing of ^{1}H-^{13}C HSQC NMR datasets. ^{1}H-^{13}C HSQC Bruker files were directly imported to Matlab R2016a using BBIO Toolbox Matlab scripts kindly provided by Bruker BioSpin GmbH, producing one (ppm1 x ppm2) matrix per spectrum. Every data matrix was then unfolded into a vector, and all vectors were merged into one matrix with as many rows as samples, and as many columns as measured ppm values. Then, data matrix was mean-centered. Before Principal Component Analysis (PCA)[2] of the yeast extract metabolomics dataset, the data were normalized using the sample factors from PQN[1] of the 1D NOESY dataset, and the same proton regions that were removed in the equivalent 1D NOESY spectra were excluded for the analysis. PQN is a convenient tool to normalize NMR spectra from time-course experiments, in which the total amount of metabolites increases over time because the studied organism is growing during the course of the experiment[3], i.e. to correct for sample size effects. Quotients used in PQN normalization are estimated from comparing the intensities relative to significant resonances of every spectrum to a reference spectrum[1].
NMR preprocessing of VOI datasets. Regions of water (δ_H = 4.41 - 5.16 ppm), methanol (δ_H = 3.30 - 3.37 ppm), chloroform (δ_H = 7.64 -7.69 ppm) and DSS (δ_H < 0.7 ppm) were removed. Resulting VOI datasets were mean-centered.

9. Principal Component Analysis

In this study, we have applied PCA to single ^{1}H-^{13}C HSQC NMR spectra and to datasets containing several ^{1}H-^{13}C HSQC NMR spectra. PCA performs an orthogonal decomposition of the analyzed spectral data sets in matrix form under the constraints of maximum variance and normalization. See refs[2] for more details about the PCA method.
In the first scenario, $\mathbf{X}_{(ppm1,\ ppm2)}$ has as many rows as ppm variables in the f1 dimension (ppm1), and as many columns as ppm variables in the f2 dimension (ppm2). On the other hand, in the second scenario, $\mathbf{X}_{(m,\ ppm1\ x\ ppm2)}$ has as many rows as investigated NMR spectra or samples, and as many columns as the total number of ppm variables in both f1 and f2 dimensions.
With PCA, data compression is performed by selecting only the principal components associated with the largest singular values[2c, 4] which will give information about the systematic variation of the data and do not describe the experimental noise, which are usually not associated to the components with the largest singular values.
In PCA, to decide the number of principal components to be considered, the singular values associated to the investigated data matrix are plotted and their sizes compared (see **Fig S3** in the Supplementary Results section for examples). It is assumed that singular values related with the relevant information of the dataset are larger than those related with random noise whose magnitude decreases slowly. Thus, the number of the largest singular values indicates the possible number of systematic variance sources (see **Fig S3**).
A more detailed explanation of the PCA method can be found elsewhere[2a, 2b].
In PCA decomposition, each new orthogonal variable explains a percentage of the variance of the initial dataset. Thus, with the analysis of the explained variance associated to every orthogonal variable, the complexity of the data can be investigated. For instance, when most of the variance (apart from noise) of a dataset containing 100 samples and 500 variables is explained by only two components, it means that most of the variance present in these 500 variables can be described by a linear combination of these two factors or components (frequently called principal components). In most of the cases, natural phenomena are driven by a limited number of physical independent sources of systematic variance apart from random experimental noise variance sources, which can be discarded.

In this study, PCA was applied directly under MATLAB R2016a environment.

Principal Component Analysis of a single ^{1}H-^{13}C HSQC NMR spectrum. For all the tested cases in this study, **X** is the original ^{1}H-^{13}C HSQC NMR spectra, containing 1,024 ppm values in f1 dimension (m=1,024), and 2,048 ppm values in f2 dimension (n=2,048).
Principal Component Analysis of multiple ^{1}H-^{13}C HSQC NMR datasets. PCA was applied directly to each mean-centered unfolded data set (see preprocessing of ^{1}H-^{13}C HSQC NMR datasets and NMR preprocessing of VOI datasets).

Results and Discussion

10. Example of a representative ^{1}H-^{13}C HSQC NMR spectrum from a metabolic yeast extract

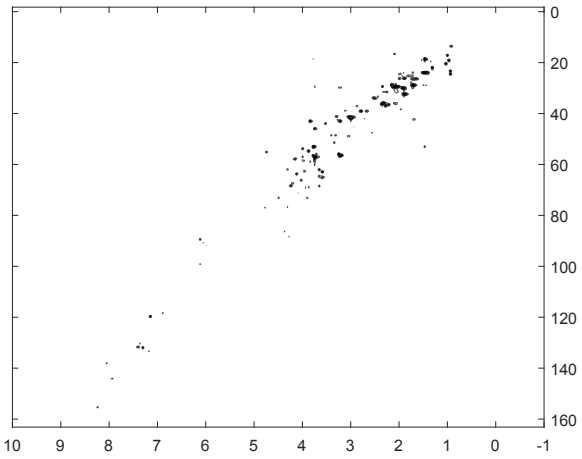

Figure S2. ^{1}H-^{13}C HSQC of a yeast metabolic extract.

11. Principal Component Analysis of a single ^{1}H-^{13}C HSQC NMR spectrum

When PCA is applied to a single 2D NMR spectral matrix, we should expect that the number of components with singular values[2c, 4] different to zero (in absence of noise) will be close to the number of detected resonances, since the only differences in intensity between the different rows should come from these resonances.

However, due to the unavoidable presence of experimental noise, as seen in **FigS3** and **TableS2**, the number of components (different to zero) in every acquired HSQC was very high (close to 500, blue lines). When the VOI-compression noise filtering algorithm was applied to the same 2D NMR spectra, the number of components decreased approximately to the number of detected resonances (red lines in **FigS3**). Some of the differences between the number of detected resonances can be explained because some of the resonances appear in the same carbon chemical shift (i.e. short-range and long-range couplings of L-glutamine and some carbons from the glucose ring).

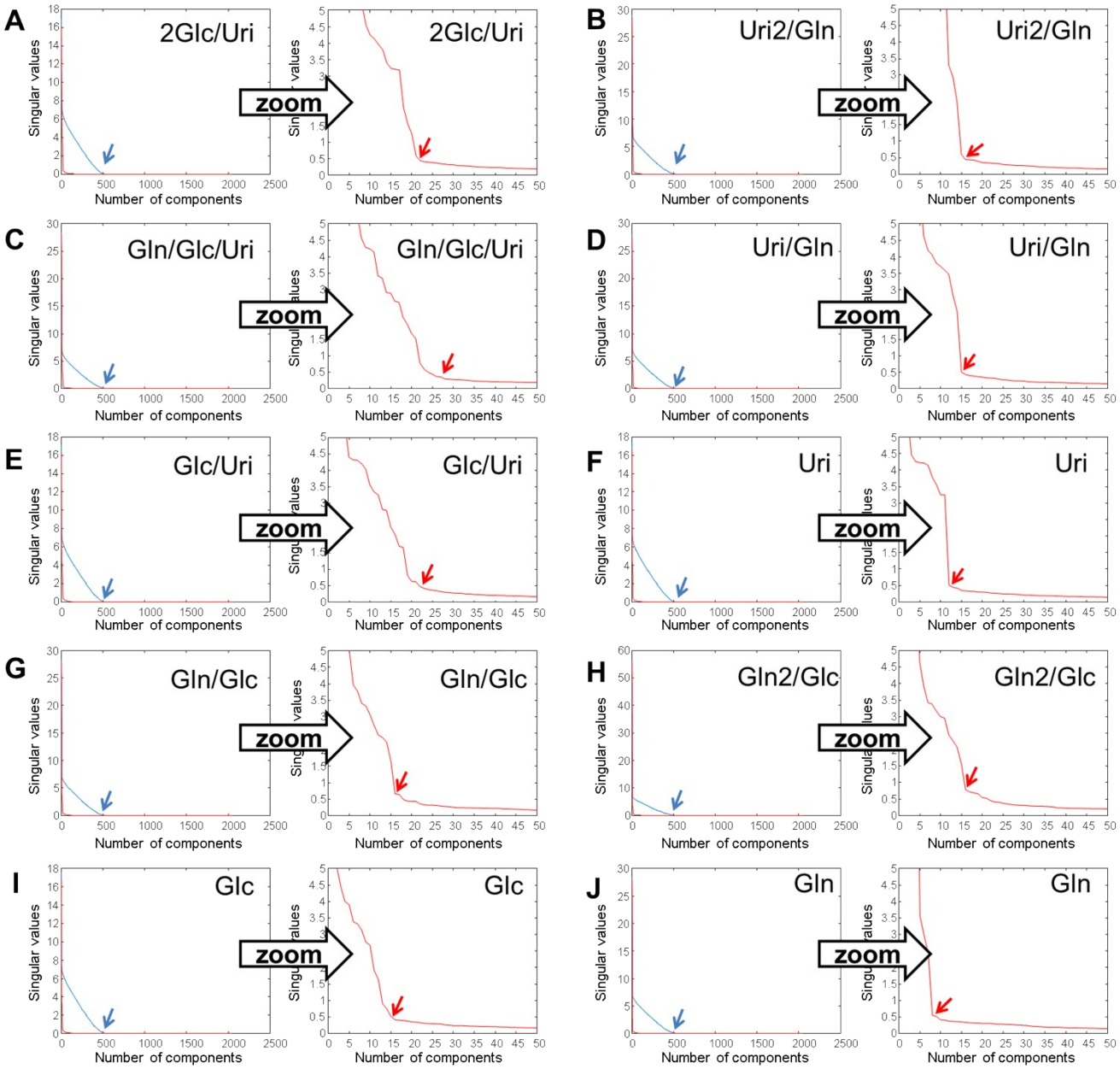

Figure S3. Plot of singular values associated to the original and VOI-processed ^{1}H-^{13}C spectra. Blue lines and arrows give the singular values of the original 2D NMR spectra, while the red lines and arrows give the singular values of the VOI-processed data.

Table S2. Number of resonances and components found in the original and VOI-processed 1H-13C spectra.

Sample name	Number of resonances*	Number of components (original spectra)	Number of components (VOI-processed spectra)
Gln	10	~500	8
2Gln/Glc	20	~500	15
Uri	12	~500	12
Gln/Uri	18	~500	16
Gln/2Uri	18	~500	16
Glc	14	~500	15
Gln/Glc	20	~500	16
Glc/Uri	22	~500	22
Gln/Glc/Uri	28	~500	28
2Glc/Uri	22	~500	22

*4 resonances were assigned to DSS, 6 to L-glutamine, 10 to D-glucose, and 8 to uridine.

12. Principal Component Analysis of multiple ^{1}H-^{13}C HSQC NMR datasets

To reduce time and computational demands, only the first m (m=10 for dataset 1, m=32 for dataset 2) components were calculated.

13. Effect on *threshold* and *minvoi* parameters

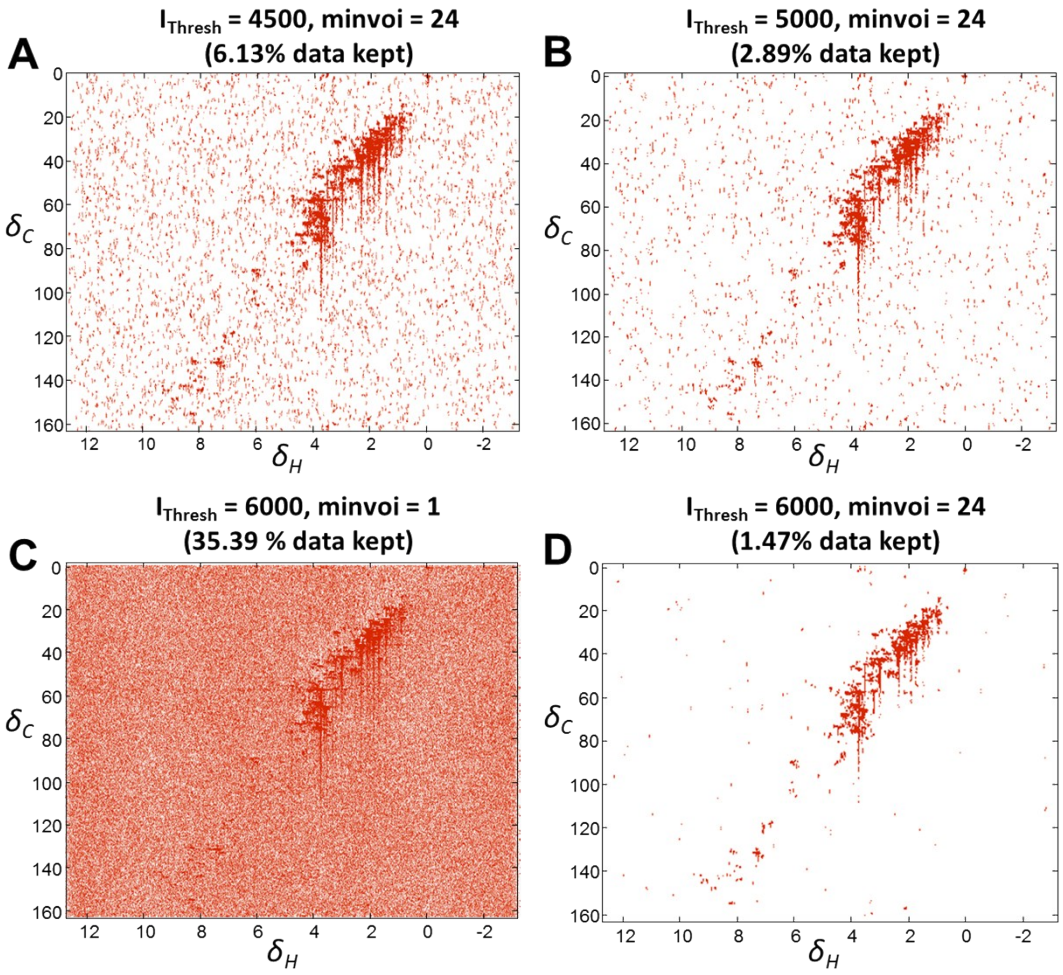

Figure S4. Selected VOIs (in red) for different applied threshold and *minvoi* levels.

When the *threshold* value was increased, a lower amount of random noise was included in the data. This is easily seen when figures A, B and D and compared.

On the other hand, not using the criterion of the minimum number of adjacent points that define a peak (*minvoi*) but maintaining the same threshold level (**Figure S4C**) results in a lower selective power. This is also explained because variables are selected when in at least one 2D NMR spectrum are higher than the *threshold*. Thus, due to the random distribution of noise, if threshold is at the same level as noise (as it is in many practical situations), the larger number of spectra analyzed, the larger number of noisy variables will be included in the data set. The only way to minimize this problem without using the *minvoi* criterion is by increasing the signal threshold value, although then this may result in the loss of some variables which are smaller than the fixed threshold level. Therefore, the simultaneous optimization of these two parameters should be better performed simultaneously.

14. VOI processing applied on 2D ^{1}H-^{1}H TOCSY experiments

Example 1: Cyclosporine sample (threshold = 100,000, minvoi = 10)

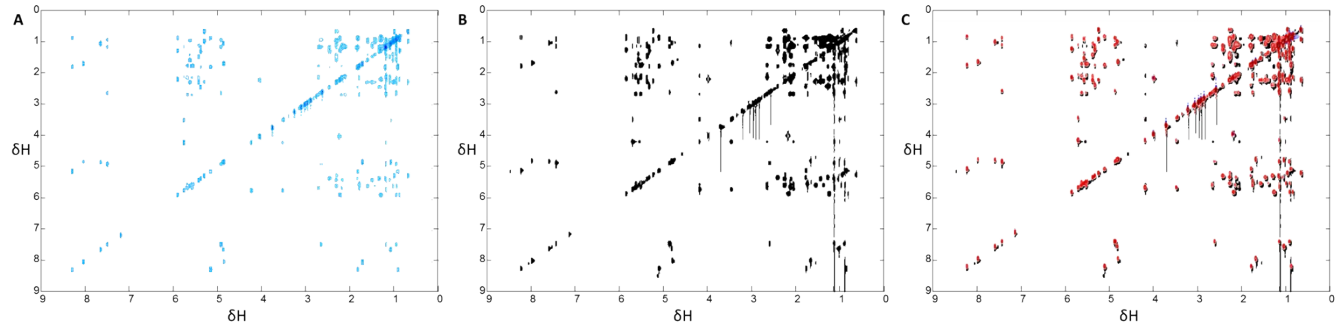

Figure S5. A) Contour plot of the reconstructed 2D ^{1}H-^{1}H TOCSY of cyclosporine sample. B) Selected VOIs (in black). C) NMR spectrum in Fig S5B overlapped with the original NMR spectrum (contour plot obtained in MestreNova NMR suite).

After application of the VOI algorithm, only 22,998 variables were selected (**Fig S5B**), which are the 2.2% of the total set of variables. In **Fig S5A**, the contour plot of the selected variables is shown. More intense peaks are colored with deep blue, whereas less intense peaks are colored with light blue. From comparing **FigS5A** and **FigS5B**, it is observed that the use of contour plots can be misleading for peak identification, as the smallest peaks may not even be plotted if not enough contour curves are used. In **FigS5C**, NMR peaks found in the reconstructed VOI-processed NMR spectrum coincide with the ones detected in the original NMR spectrum.

Example 2: AcCNPNFDLEC sample (threshold = 14,000, minvoi = 40)

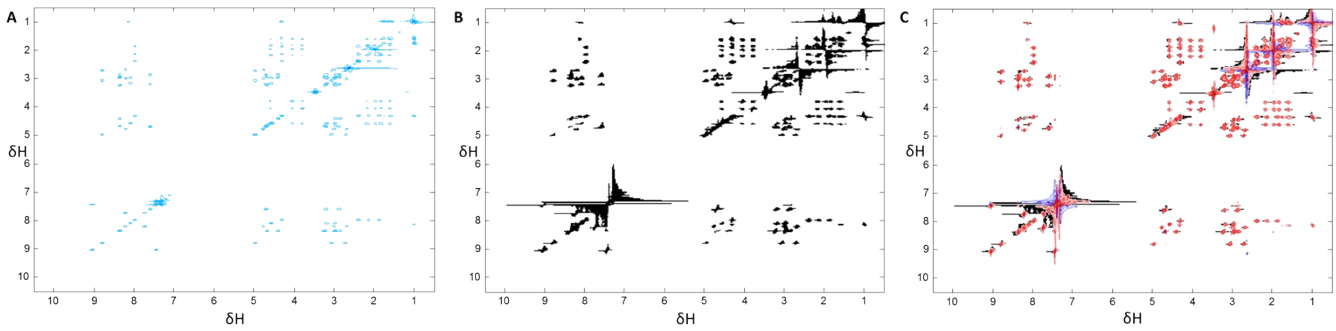

Figure S6. A) Contour plot of the reconstructed 2D ^{1}H-^{1}H TOCSY of AcCNPNFDLEC sample. B) Selected VOIs (in black). C) NMR spectrum in Fig S6B overlapped with the original NMR spectrum (contour plot obtained in MestreNova NMR suite).

After application of VOI algorithm, only 29,833 variables were selected (**Fig S6B**), which corresponds to 2.8% of the total set of variables.

15. VOI processing applied on 2D ^{1}H-^{15}N HSQC experiments

Example 3: Ubiquitin sample (threshold = 3,000, minvoi = 40)

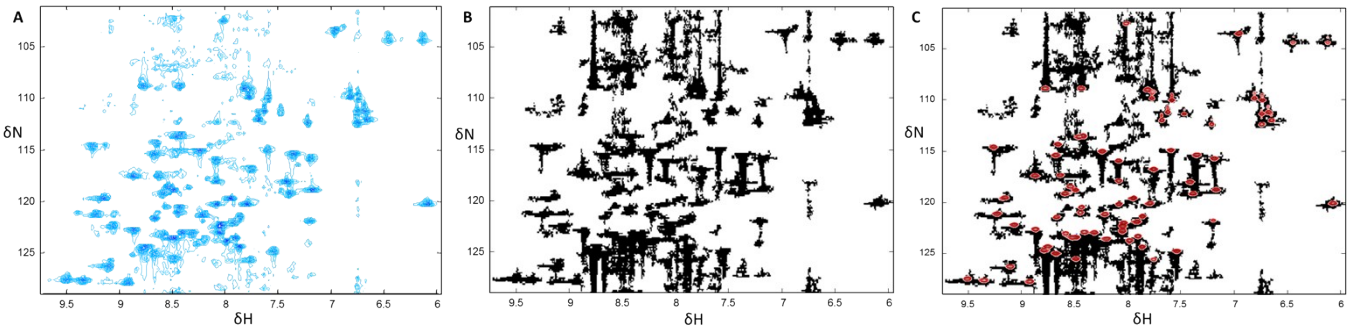

Figure S7. A) Contour plot of the reconstructed 2D ^{1}H-^{15}N HSQC of ubiquitin sample. B) Selected VOIs (in black). C) Overlapped NMR spectrum of Fig S7B with the original NMR spectra (contour plot obtained in MestreNova NMR suite).

After application of VOI algorithm, only 18,660 variables were selected (**Fig S7B**), which corresponds to 14.6 % of the total set of variables. In this example, the number of selected variables was higher than in the previous examples because most peaks present tales along f1 that were also selected. With this example, it is proven that even for not very well resolved NMR spectra in both

dimensions (the HSQC used here has 213 data points in f1 and 602 data points in f2) the analysis can be performed properly by the proposed VOI approach.

16. VOI processing applied on 2D ^{1}H-^{1}H ROESY experiments

Example 4: AcCNPNFDLEC sample

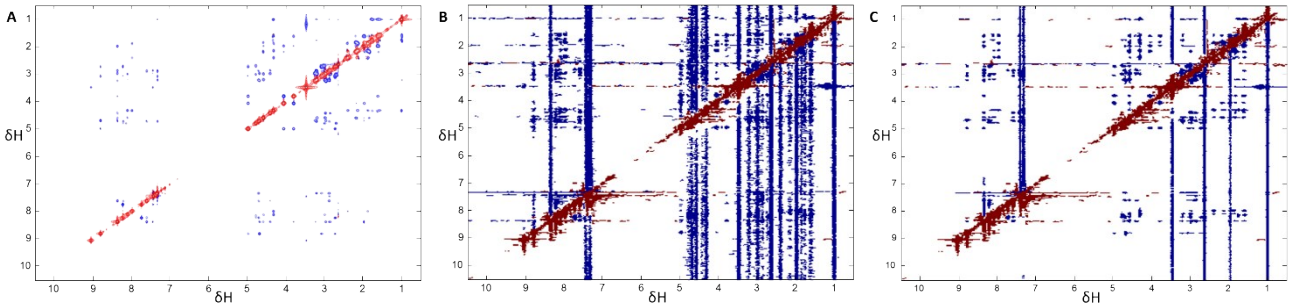

Figure S8. A) Contour plot of the original 2D ^{1}H-^{1}H ROESY of AcCNPNFDLEC sample (in MestreNova NMR suite). B) Contour plot of the selected VOIs in the analysis of the 2D ^{1}H-^{1}H ROESY of AcCNPNFDLEC sample (in MATLAB with a threshold_positive = 1,400; threshold_negative = -1,400; minvoi = 25). C) Contour plot of the selected VOIs in the analysis of the 2D ^{1}H-^{1}H ROESY of AcCNPNFDLEC sample (in MATLAB threshold_positive = 1,400; threshold_negative = -4,000; minvoi = 25). Blue color is associated to negative intensity values, and red color is associated to positive intensity values.

The VOI-processing strategy can be also applied to the phase-sensitive 2D NMR spectra such as in 2D ^{1}H-^{1}H ROESY experiments. As stated before in section 7, to deal with positive and negative peaks, two threshold levels were used.

In the example 4 (**Figure S8**), a 2D ^{1}H-^{1}H ROESY experiment was processed using the strategy based on VOIs. When the two threshold levels were 1,400 and -1,400, corresponding approximately to the 50% of the maximum and minimum noise levels, respectively, 161,396 variables were selected (15.4% of the total set of variables). In **FigS8B**, it is observed that most of the selected variables did not correspond to noise, but to structured negative bands found mostly along f1. These structures were consequence of the NMR pulse sequence used. In order to remove most of these meaningless data values with systematic (not random) information, negative threshold can be set up a bit lower. When this parameter was fixed at -4,000, the number of selected variables decreased down to 72,790, which corresponds to the 7.9% of the total set of original variables.

17. VOI processing applied on a 3D HNCO experiment

Example 5: Protein sample (threshold = 150,000, *minvoi* = 40)

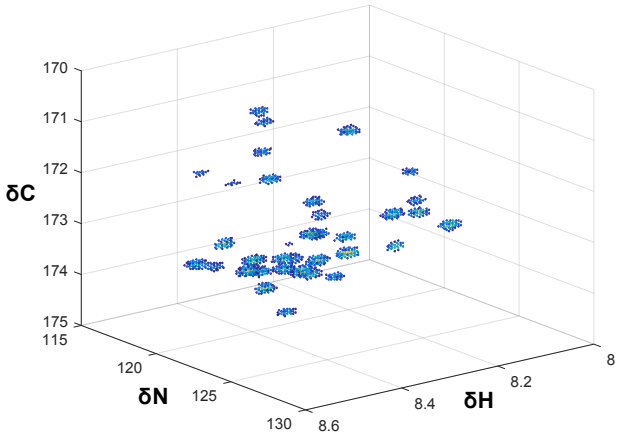

Figure S9. 3D plot of a highlighted region of the filtered HNCO NMR spectrum, where intensity is represented by a color scale.

The VOI-processing strategy can be also applied to 3D NMR spectra such as in 3D HNCO NMR experiments.

In the example 5 (**Figure S9**), a 3D HNCO NMR experiment was processed using the strategy based on VOIs. The input 3D NMR spectra consisted on a cubic dataset with dimensions of 1,024 x 256 x 256 (after Topspin processing of 2048/72/128 acquired TD points, applying zero-filling and strip transform with STSI=1024), giving a total of 67,108,864 variables that occupy 256 MB (file 3rrr). These dimensions correspond to 1,024 δ_H values (from 4.77 to 12.89 ppm), 256 δ_N values (from 100.09 to 136.09 ppm) and 256 δ_C values (from 165 to 181 ppm), respectively.

In the 2D VOI-processing strategy, variables are searched on the 8 different positions contained in the X-Y plane (upper-left, up, upper-right, left, right, lower-left, low, lower-right). For the VOI-processing strategy extended to 3D NMR data, 26 positions are considered instead (8 positions for the same X-Y plane than the investigated variable (with z=0), and 9 positions for the X-Y planes with z=+1 and z=-1).

When the threshold level was set to 150,000 and the *minvoi* was set to 40, 125,678 variables were selected (0.19% of the total set of variables). These variables were grouped in 61 clusters (individual or overlapped 3D resonances). In **Figure S9**, a highlighted region of the reconstructed 3D NMR spectrum, with 89 δ_H values, 61 δ_N values and 101 δ_C values, is shown.
In **Figure S10**, five ^{1}H-^{13}C slices at different ^{15}N chemical shift from **Figure S9** are given.

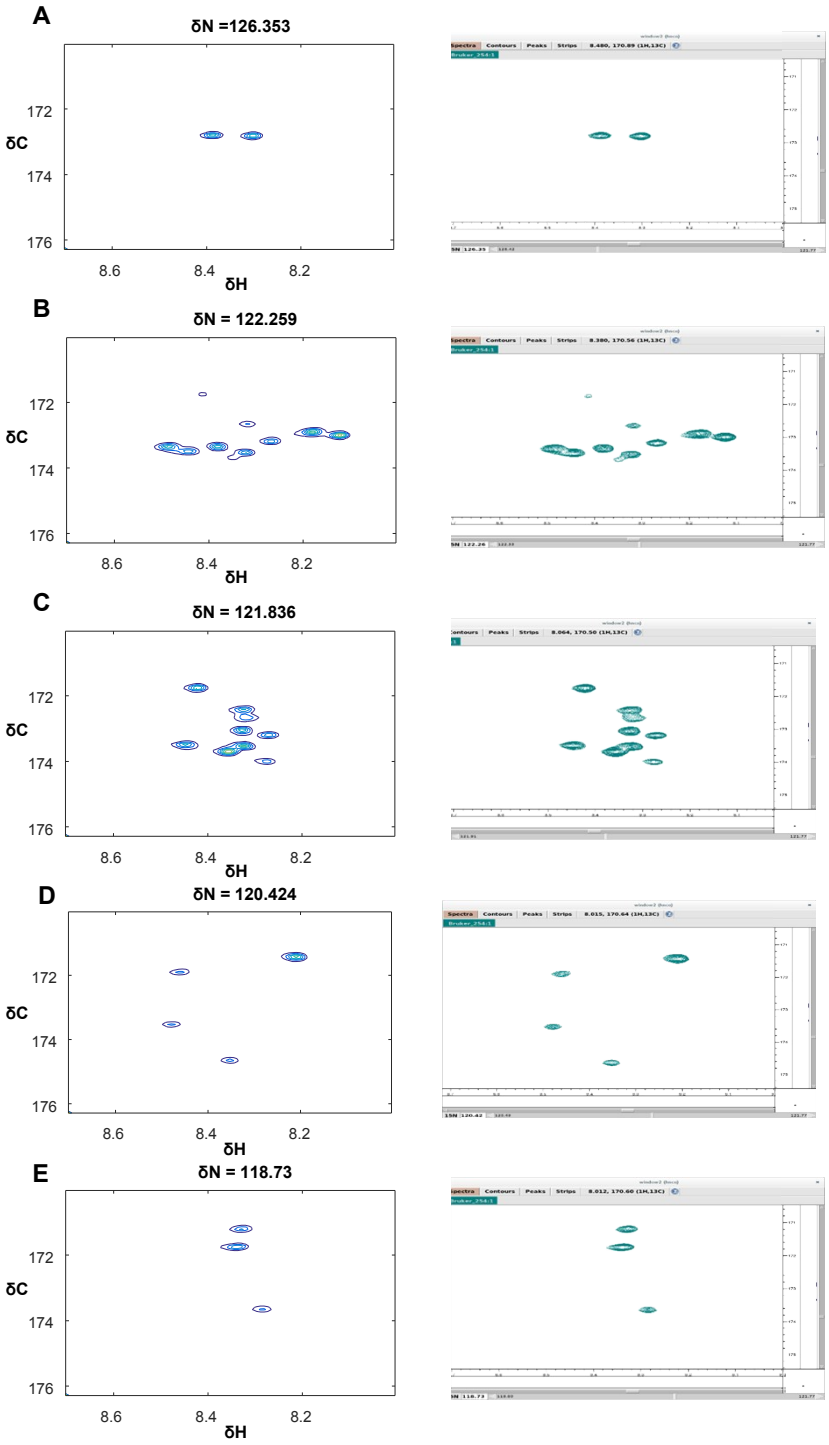

Figure S10. Selected ^{1}H-^{13}C slices at different 15N chemical shifts (A-E) from the 3D region highlighted in **Figure S9.** The left spectra were the reconstructed spectra after VOI filtering, while the right ones are the original raw data viewed in CcpNmr software.

References

[1] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, *Anal. Chem.* **2006**, *78*, 4281-4290.

[2] aI. Jolliffe, in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, **2014**; bR. Bro, A. K. Smilde, *Anal. Methods* **2014**, *6*, 2812-2831; cH. Abdi, L. J. Williams, *Wiley Interdisciplinary Reviews: Computational Statistics* **2010**, *2*, 433-459.

[3] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, *Scientific Reports* **2016**, *6*, 30982.

[4] H. Abdi, in *Encyclopedia of measurement and statistics* (Ed.: N. J. Salkind), SAGE Publications, **2007**, pp. 907-912.