Journal of proteome research

Article

Tackling the Peak Overlap Issue in NMR Metabolomic Studies: 1D Projected Correlation Traces from Statistical Correlation Analysis on Nontilted 2D H NMR *J*-Resolved Spectra

Andrés Charris-Molina, Gabriel Riquelme, Paula Burdisso, and Pablo A. Hoijemberg

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.9b00093 • Publication Date (Web): 27 Mar 2019

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Tackling the Peak Overlap Issue in NMR Metabolomic Studies: 1D Projected Correlation Traces from Statistical Correlation Analysis on Nontilted 2D ^{1}H NMR *J*-Resolved Spectra

Andrés Charris-Molina,[†,‡,⊥] Gabriel Riquelme,[†,‡,⊥] Paula Burdisso,[§] and Pablo A. Hoijemberg*[,†,∥]

[†] CIBION-CONICET, Centro de Investigaciones en Bionanociencias, NMR Group, Polo Científico Tecnológico, Ciudad Autónoma de Buenos Aires C1425FQD, Argentina

[‡] Universidad de Buenos Aires, Departamento de Química Inorgánica Analítica y Química Física, Facultad de Ciencias Exactas y Naturales, Ciudad Universitaria, Buenos Aires C1428EGA, Argentina

[§] Instituto de Biología Molecular y Celular de Rosario (IBR-CONICET), Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario and Plataforma Argentina de Biología Estructural y Metabolómica (PLABEM), Rosario, Santa Fe, Argentina

[∥] ECyT-UNSAM, 25 de Mayo y Francia, San Martín, Buenos Aires B1650HMP, Argentina

ABSTRACT: The identification of metabolites in complex biological matrices is a challenging task in 1D ^{1}H NMR based metabolomic studies. Statistical TOtal Correlation Spectroscopy (STOCSY) has emerged for aiding the structural elucidation by revealing the peaks that present high correlation to a driver peak of interest (which would likely belong to the same molecule). However, in these studies the signals from metabolites are normally present as a mixture of overlapping resonances, limiting the performance of STOCSY. 2D ^{1}H homonuclear *J*-resolved spectra (JRES), in its usual tilted and symmetrized processed form, were projected and STOCSY was applied on these 1D projections (p-JRES-STOCSY) as an alternative to avoid the overlap issue, but this approach suffers in cases where the signals are very close. In addition, STOCSY was applied to JRES spectra (also tilted) to identify correlated multiplets, although the overlap issue in itself was not addressed directly and the subsequent search in databases is complicated in cases of higher order coupling. With these limitations in mind, in the present work we propose a new methodology based on the application of STOCSY on a set of nontilted JRES spectra, detecting peaks that would overlap in 1D spectra of the same sample set. COrrelation COmparison Analysis for Peak Overlap Detection (COCOA-POD) is able to reconstruct projected 1D STOCSY traces that result in more suitable database queries, as all peaks are summed at their f2 resonances instead of the resonance corresponding to the multiplet center in the tilted JRES (the peak dispersion and resolution enhancement gained are not sacrificed by the projection). Besides improving database queries with better peak lists obtained from the projections of the 2D STOCSY analysis, the overlap region is examined and the multiplet itself is analyzed from the correlation trace at 45° to obtain a cleaner multiplet profile, free from contributions from uncorrelated neighboring peaks.

KEYWORDS: NMR, STOCSY, *J*-resolved, metabolomics, metabolite, identification, database, correlation matrix, overlap, complex mixture

INTRODUCTION

The identification of compounds in complex mixtures is critical for any metabolomics study, which aims at identifying differences in concentrations of metabolites in distinct groups (populations, species, cohorts, etc.).[1,2] These metabolic differences result in a biological interpretation of the changes between (or among) the metabolic profiles of the analyzed samples, and reducing the amount needed for a proper identification of the metabolites is key in decreasing the time of one of the bottleneck steps within a metabolomics study.

Statistical correlation analysis, STOCSY,[3] has been for over one decade one of the tools of choice to aid in the identification of compounds in NMR based metabolomics studies by revealing the peaks that present high correlation to a driver peak of interest, namely one that is present in loadings or coefficient plots and is responsible for classification after a multivariate data analysis (MVDA) over the spectral data. The list of peaks showing high correlation with the driver peak, including the driver peak itself, can be used for a query at any of the several databases available, like HMDB, BMRB, etc.[4,5]

Several adaptations were introduced to STOCSY to improve its performance, to combine with other nuclei or analytical platforms and to extract biological information from perturbed metabolic pathways,[6] as well as alternatives (like ratio analysis NMR spectroscopy[7]). However, STOCSY presents limitations where its performance is diminished, mainly in cases of peak misalignment, weak peaks and specially with peak overlap.[3] The peak misalignment issue, mostly due to pH sensitive species whose peaks shift depending on the pH of the sample, can be corrected with peak alignment tools during the stage of data processing,[8] or can be ideally avoided or controlled by using a buffer solution during sample preparation (with or without pH adjustment).[9] The weakness of the peaks can only be improved with a longer experiment time, as more scans are expected to increase the signal to noise ratio. Finally, the peak overlap issue may be partially avoided by adjusting sample preparation. For example, a simple sample treatment procedure can be used to deal with peak overlap due to carbohydrates signals, by treating the sample with $NaIO_4$ and eliminating their peaks in the 3.2 to 4.5 ppm region.[10] Otherwise, dealing with peak overlap requires further experiments besides the standard 1D ^1H analysis, essentially 2D experiments or relaxation filtered 1D experiments and/or data analysis tools applied to the acquired spectra.

Only a few methodologies were reported that can address the peak overlap issue to some extent. STORM,[11] developed at Imperial College London, selects subsets of spectra that are similar in a region of interest by matching to a reference spectrum, and was mainly developed to extract information from weak peaks and discriminate them from the noise. If applied to a region of interest with overlapping peaks and selecting a subset where the overlap is not present, it could render a subset in which STOCSY is expected to perform better. Nonetheless, its usability relies on the existence of a subset of spectra lacking the peak overlap in the region of interest.

A recent approach, POD-CAST,[12] was applied on STOCSY analysis of 1D spectra to detect peak overlap regions by identifying the decrease in correlation caused by the peak overlap (nonstructural correlation), when compared to the high correlation from peaks in the same molecule (structural correlation). It performs also a comparison of the correlation traces by clustering analysis to create peak lists that result in better database queries, after overlapping peaks are added to the lists missing them. However, it fails when the overlap is such that no contribution is observed by one or more of the overlapped

peaks, and it also fails to properly distinguish the origin of medium intensity correlations, in other words, biological correlation contributions from overlapping peaks. This final analysis can thus be performed with user intervention, but not fully automated.

The peak overlap issue can also be addressed by the use of two-dimensional NMR experiments, which spread the overlap in the ^{1}H spectra by adding an additional (indirect) dimension in which the overlapping signals have different resonance frequencies (for example COSY, HSQC, TOCSY, JRES, etc.). Hence, one can find cross peaks in a HSQC spectrum for signals at the same ^{1}H chemical shift, but at different ^{13}C chemical shifts, as exemplified in the use of statistical correlation spectroscopy applied on HSQC data sets.[13,14] There are evident drawbacks in cases like this: the databases are more limited in their content of HSQC data, and the added information present in the ^{1}H spectra, integration and coupling, is not available to aid in structure elucidation. Two-dimensional experiments normally require longer acquisition times than 1D ^{1}H experiments, and 2D ^{1}H homonuclear *J*-resolved (JRES) experiments present higher sensitivity than HSQC experiments, being the JRES the one mostly used for the standard metabolomics experiments in the whole sample set. JRES experiments also present a good cost to benefit compromise due to the coupling information it provides in the indirect dimension, which aids in the compound identification process.

An older approach to deal with the peak overlap issue is the use of JRES spectra, which when processed with the usual 45° tilt, symmetrization and with either the sum or skyline projections produces 1D spectra that resemble homonuclear decoupled ^{1}H spectra. In these projections (p-JRES), each signal is ideally a singlet, centered at the center frequency of the corresponding multiplet. This results in a spread of the peaks (there are less peaks than in the original 1D spectrum) with the condensation of the multiplets into a single peak, and STOCSY can be performed on these p-JRES spectra.[15] An analogous analysis can be performed directly on spectra obtained with pure shift experiments,[16] or from the projections of the PSYCHEDELIC experiment after the usual 45° tilt processing step.[17] Nonetheless, the use of the p-JRES spectra suffers from multiplet overlap and is easy to employ only if first order coupling occurs, as it is a complex task to generate a proper list of peaks for a database search from multiplets exhibiting strong coupling.

A better use of the JRES spectral data was obtained with the 2D extension of STORM published recently, RED-STORM,[18] and the application of STORM in 2D experiments, referred to as STORM$_2$.[11] With the same purpose of STORM, the spread of peaks in the indirect dimension containing the *J*-coupling information helps RED-STORM in the identification of peaks with structural correlation that would otherwise remain obscured by overlapping peaks with higher contribution, which STORM in 1D would have missed. Still, like STORM, RED-STORM is based on the creation of a subset for reference matching and an algorithm comprised of several steps after the simple first statistical correlation calculation. In

addition, it works with the JRES spectra 45° tilted and symmetrized, which has an increased risk of incorporating artifacts due to strong coupling. Akin to the use of STOCSY on the p-JRES, the results of $STORM_2$ and RED-STORM are clearly useful for cases of first order coupling, as in the two molecules identified with RED-STORM. Nonetheless, it seems to be complicated for interpretation when there are multiplets with second order coupling, with the mentioned difficulty in the construction of peak lists for database query, and the variable selection employed in RED-STORM.

Lastly, either approach employing JRES spectra relies on the availability of adequate searchable databases, either for multiplet chemical shift values, for p-JRES, or for the whole multiplet information (chemical shift, multiplicity and coupling constants), for RED-STORM or $STORM_2$ for example. SpinCouple[19] aims to cover all of this multiplet information for JRES spectra, and was constructed on JRES spectra acquired by the authors (more than 600 compounds listed to date) and compared to data from the Birmingham Metabolite Library.[20] Despite the usefulness of the information obtained as output, neither the old nor the new versions of SpinCouple available at the RIKEN Environmental Metabolic Analysis Research Team website provide the results as hits or scoring per compound, like HMDB[4] or COLMAR[21] do, rendering the analysis of the query output truly cumbersome. As the Birmingham Metabolite Library (containing spectra from 208 selected metabolites) only allows to search by compound name, it seems more reliable to search for data obtained from 2D experiments on 1D databases allowing for faster and high throughput queries.

In the present work we propose a methodology that aims to detect the peaks spread in a JRES spectra that would overlap in 1D spectra of the same sample set with the purpose of obtaining a better identification through statistical correlation analysis on two dimensions, projecting the result to mimic a 1D STOCSY trace for database query and comparing to what could be obtained directly through STOCSY applied on the 1D spectra set. The methodology, named COrrelation COmparison Analysis for Peak Overlap Detection (COCOA-POD), utilizes the nontilted JRES spectra (ntJRES, 2D), whose projections (p-ntJRES, 1D) resemble CPMG spectra, instead of the usual p-JRES spectra, 1D, which are projections after a 45° tilt and symmetrization of the JRES spectra, employed in JRES-STOCSY (referred to as (p-JRES)-STOCSY in this work).[15] This permits the independent analysis of all the peaks from a whole multiplet, as an alternative to its tilted sum projection, as peaks from a multiplet can be overlapping with peaks from other multiplets in a 1D ^1H spectrum.

METHODS

Simulated spectra set. Raw free induction decay files corresponding to 2D JRES NMR experiments were downloaded from the Birmingham Metabolite Library[20] for solutions of the levorotatory forms of alanine, glutamic acid, glutamine and ornithine. Spectra were processed with the following parameters: zero filling to generate a 16k

by 128 points data matrix for each spectrum (n = 165), a sine bell window and an exponential window (LB = 0.5 Hz) were applied to f2, while a sine bell window was used on f1.[22] Slight chemical shift adjustments were performed on each spectral referencing step to ensure a great extent of overlap at 3.78 ppm of the following signals for each amino acid: second most upfield peak for alanine, most downfield peak for glutamic acid, triplet center peak for glutamine and most upfield peak for ornithine. These shifts were at most of 0.0075 ppm, which did not alter the query results in HMDB for any of the amino acids. All individual amino acid spectra were normalized to an equal integral value for the α ^{1}H signals around 3.75-3.82 ppm. Linear combinations of these basis spectra were created using randomly generated coefficients, with values constrained between 1.0 and 1.4, to obtain 165 linearly independent JRES spectra. For most of the work the spectra were nontilted, ntJRES, and were only tilted 45° to obtain the "uncoupled" projections for JRES-STOCSY,[15] p-JRES spectra.

Blood serum set, sample collection. Blood samples from 150 healthy volunteers from the central region of Argentina were collected. Two different collection methods were used, in one case blood extraction was carried out with a syringe (63 samples) and in the other it was carried out using vacutainers SST BD Bioscience (87 samples). After collection covered test tubes were left in a standing position for 30 minutes and centrifuged at 20^{0}C, 1400 g for 10 min. Finally, 1 mL supernatant was removed into labelled screw top cryo-vial for storage at -80 °C until analyzed. Informed consent was signed by all volunteers.

Blood serum set, sample preparation. Samples were thoroughly thawed at room temperature. 300 µL of serum were mixed with 300 µL of 0.075 M phosphate buffer pH 7.4 containing 0.08% (w/v) sodium 3-trimethylsilyl-(2,2,3,3-$^2H_4$)-1-propionate (TSP); 0.04% (w/w) NaN_3 and 10% (v/v) D_2O. Samples were then centrifuged at 12000 g for 10 minutes at 4°C and 500 µL of the supernatant were transferred to an NMR tube. A pooled quality control sample (QC) was prepared by mixing equal volumes (50 µl) from all 150 samples.

NMR Data Acquisition. ^{1}H spectra were obtained at 310 K using a Bruker Avance 600 MHz NMR spectrometer (Bruker Biospin, Rheinstetten, Germany) equipped with a 5-mm TXI probe. One-dimensional ^{1}H NMR spectra of serum were acquired using a Carr-Purcell-Meiboom-Gill (CPMG) spin echo sequence (cpmgpr1d)[23,24] with water presaturation.[25] The mixing time was set to 10 ms, the data acquisition period to 2.73 s and the relaxation delay to 4 s. ^{1}H NMR spectra were acquired using 4 dummy scans and 64 scans, with 64k time domain points and a spectral window of 20 ppm. FIDs were multiplied by an exponential weighting function (LB = 0.3 Hz).

J-resolved pulse experiments[25] (jresgpprqf) were acquired with suppression of the water resonance -RD–90°–t1–180°–t1–ACQ, where t1 is an incremented time period, RD is 2 s and 180° represents a 180° RF pulse, ACQ is 0.41 s. *J*-resolved spectra were acquired using 16 dummy scans and 1 scan, 8k points with spectral window of 16.7

ppm for f2 and 40 increments with spectral window of 78 Hz for f1. Continuous wave irradiation was applied at the water resonance frequency using a 25 Hz RF during the RD. One experiment from a quality control sample was acquired every 10 samples from volunteers. Additional JRES spectra of a representative sample, a quality control, were acquired using 1, 2, 4, 8, 16 and 32 scans.

NMR Data Processing. JRES spectra was zero filled to generate a 16k by 128 points data matrix for each spectrum (n = 165). A sine bell window and an exponential window (LB = 0.5 Hz) were applied to f2 and a sine bell window to f1,[22] followed by a Fourier transform in both dimensions and conversion to magnitude mode. JRES spectra were baseline corrected to remove the residual water signal. The spectra were not tilted prior to the statistical analysis. The direct dimension of the spectra was referenced to the glucose peak at δ = 5.233 ppm.[25] In the set of experiments with 1 scan there were some spectra with peaks likely belonging to formate (singlet at 8.44 ppm) and tyrosine (two doublets, at about 7.18 and 6.88 ppm), at intensity levels at or below the detection limit. However, there were no peaks outside of the 0-6 ppm region in the mean ntJRES spectrum, and the downfield and upfield extremes were consequently removed for the statistical analysis step. The resulting spectra of all samples were exported to a text file prior to data analysis. The suppressed water peak region (4.5-5.0 ppm) was set to zero intensity and the residual water peak intensity left beyond that region was baseline corrected with polynomial functions for each f2 trace. For the analysis with increasing number of scans the region downfield of 6 ppm was considered and not removed, as increasingly more peaks appear in the aromatic region.

Statistical analysis. All preprocessed spectra were imported into MATLAB R2014b (Mathworks, Natick, MA) using in-house written functions. JRES spectra were normalized to constant volume and stacked into a 3D array (f1 by f2 by *n*). Data points were interpolated in f2 to produce a data matrix with equal resolution in f2 and f1, which allowed for an easy extraction of the multiplet trace at 45°, and a smooth shear transformation to produce the tilted spectra. This data array was used to calculate mean ntJRES and mean p-ntJRES spectra, for the latter summing all values along f1 and grouped in a new *n* by f2 2D array.

Selected (f2, f1) pairs or f2 values (local maxima in regions of interest) were used as driver peaks to generate STOCSY[3] matrices or traces from ntJRES, p-JRES and p-ntJRES data. In the case of ntJRES-STOCSY, the 3D data array was reshaped into a 2D *n* by (f1 × f2) matrix and this new array was used to calculate the STOCSY traces in a similar way to a one dimensional dataset. The resulting traces were shaped back into a pseudo 2D spectrum composed of covariance and correlation matrices. As the objective of the statistical correlation analysis was to find structural correlations, the correlation color-coding for the STOCSY contours or traces was displayed for positive values, ignoring negative correlations that may appear.

The projections of the ntJRES-STOCSY matrices, p-(ntJRES-STOCSY), were calculated for all 2D STOCSY matrices summing all covariance values along f1 that possessed a correlation value greater than a given

threshold (default value was 0.5), thus reducing the probability of including in the projection peaks from non-structural correlation and keeping peaks with high probability of being from structural correlation.[26] Correlation color-coding in the covariance projected trace was based on the maximum values for correlation along each f1 trace. Contour levels for peaks with correlation values above the threshold were considered only above a covariance noise level, to avoid inclusion of incidental high correlation peaks with covariance at the noise level of the ntJRES-STOCSY spectra. For 1D STOCSY, the statistical analysis was performed directly on the p-ntJRES or p-JRES traces and represented as usual for a STOCSY trace. The resulting (p-ntJRES)-STOCSY traces are compared to the p-(ntJRES-STOCSY) traces, to account for differences and identify overlap instances.

Multiple testing for the multiple correlations calculated in the 2D STOCSY analysis on the ntJRES spectra was accounted for by performing a test for zero population correlation with an alternative hypothesis for the correlation being different from zero.[27] With $\alpha = 0.05$ and a Bonferroni correction by the total number of elements in the 2D matrix, a limiting correlation coefficient, r_{lim}, was calculated ($r_{lim} = 0.41$). Below this r_{lim} the difference between any r and zero is not statistically significant and it allowed us to disregard all correlation coefficients below it (more detail is included in Section 1 of the Supporting Information). Any correlation threshold chosen for the sum projections above the calculated r_{lim} ensures that the projected peaks arise from correlations significantly different from zero.

The COCOA-POD figure comprises a set of subplots comparing p-(ntJRES-STOCSY) and (p-ntJRES)-STOCSY pseudospectra. Subplot A presents the contour plot for the mean ntJRES spectrum, with a mark over the driver peak together with the color-coded ntJRES-STOCSY contour plot superimposed. The right section of subplot A, a horizontal zoom for the region around the driver peak, clarifies the panorama with regard to the presence (or absence) of peak overlap for the driver peak at f2. The f1 trace represented by the green dashed line, and the contribution from neighboring multiplets to the driver peak multiplet trace at 45°, violet dashed line, obviously intersect at the driver peak, and will be analyzed in a separate plot together with the multiplet analysis for clarity purposes. Subplot B reveals the projection of the STOCSY plot in subplot A, p-(ntJRES-STOCSY), with the zoom at the driver peak region revealing the multiplet structure. Finally, subplot C displays the (p-ntJRES)-STOCSY trace with driver peak at the f2 value in subplot A, mimicking the STOCSY analysis over a CPMG spectral set, superimposed over the mean p-ntJRES spectrum. The horizontal zoom for subplot C can be compared directly with that for subplot B, for coincidence or dissimilarity, the latter indicating also overlap at f2, evidenced in the magnification of subplot A as well. Scheme S1 in the Supporting Information summarizes the whole procedure to obtain the data in subplots A, B and C, and the algorithm

is presented as pseudocode in Section 1 of the Supporting Information as well.

For the overlap trace, violet dashed line in the zoom of subplot A, the mean ntJRES spectrum values at f2 for the driver peak are plotted, with the covariance trace from the ntJRES-STOCSY matrix at the same f2 superimposed (both intensity and covariance normalized to the same value at the f1 of the driver peak). In addition, an estimation of the driver peak area relative to the total area is performed by simple integration under the curve. A peak with no overlap at f2 will represent 100% of the area, with the mean ntJRES and ntJRES-STOCSY traces overlapping. A peak with any degree of overlap will present other peaks besides that of the driver peak. While usually this is evidenced with the other peaks lacking correlation, it might happen that there is correlation to the other peak (or peaks) if they belong to the same molecule (excluding cases of biological correlation from this analysis).

Metabolomics studies, as any case of complex mixtures, usually suffer from either partial or total peak overlap, between (or among) different multiplets, which hinders the determination of coupling constants. Multiplet structure of the 45° traces containing the driver peak from the ntJRES-STOCSY spectra are equivalent to the f1 traces of the tJRES spectra at the multiplet center chemical shift. This ntJRES-STOCSY 45° trace is presented superimposed to that of the mean ntJRES spectrum at 45°, to evaluate the possible contamination of the multiplet with peaks from neighboring multiplets. This gives cleaner spectra that can be analyzed with better results, with overlapping peaks having been eliminated (or neglected based on low correlation values). The 45° ntJRES-STOCSY traces were analyzed using a modified version of the algorithm designed by Hoye et al,[28] which harness the symmetries and the peak intensity ratios present in a first order multiplet to estimate the coupling constants in the system, assuming that the multiplet is pure. Due to the characteristic phase twist distortion of the JRES peaks and possible intensity distortions due to leaning effects in coupled nuclei with close chemical shift values, the algorithm was allowed some tolerance for the estimation of the number of couplings present, as the final intensities of the peaks did not always have the required integer values for the peak to peak intensity ratios. Based on the resolution in the f1 dimension there is an uncertainty on the estimated *J* coupling values of 0.6 Hz.

Database query. Lists of chemical shifts were obtained for 1D ^{1}H database query at HMDB (using 0.02 ppm tolerance) by executing a peak picking routine over the p-(ntJRES-STOCSY) traces, subplot B of COCOA-POD, and excluding the peaks having correlation coefficients below 0.85. This value was chosen to be slightly below the recommendation of $\theta = 0.89$ for a positive predictive value of 0.9, according to Couto Alves et al.,[26] to allow some peaks right above the detection limit that present weaker correlation coefficients to be included. The database query were conducted with these lists, noting the identification for the first hit, its Jaccard Index (JI) and Match Ratio (MR). The same peak picking routine was applied to the (p-ntJRES)-STOCSY traces (subplot C), where different

results are expected in cases of peak overlap. Due to the known overlap in the the α ^{1}H region for the simulated set, an additional lower threshold value of 0.65 was selected, to include more peaks whose correlation coefficients might have been reduced due to the overlap.[26]

Some peaks in the simulated set corresponding to residual solvent signals (present in the original database raw files[20]) where excluded, clearly identified due to their low intensity relative to the amino acids peaks. Likewise, in the blood serum set some peaks identified as consequence of biological correlation were excluded from the search to improve the database query results (an initial search with all the peaks was performed, and the nonmatching peaks analyzed separately). For example, in the 3-hydroxybutyric acid spectrum there are also signals from the other ketone bodies, acetone and acetoacetate. When appropriate, additional information related to the multiplets in the p-(ntJRES-STOCSY) traces was compared to the multiplet detailed information in the HMDB[4] to increase the level of confidence of compound annotation[29] for the metabolites in the blood serum set.

RESULTS AND DISCUSSION

COCOA-POD in simulated spectra set. The essence of this work comes from avoiding tilting the 2D JRES spectra at all. The benefit of not tilting the spectrum is that the projection of a ntJRES spectrum, p-ntJRES, is quite similar to a CPMG spectrum for the same sample, even for multiplets exhibiting strong coupling, as discussed below for ornithine as an example. A similar approach was reported by Shapiro and coworkers for the study of resin bound molecules derived from combinatorial chemistry using magic angle spinning NMR,[30] where the projection of the nontilted JRES spectrum produced a spectrum with higher resolution than a spin echo experiment and at the same time eliminated the broad signals from the polymer itself. To the best of our knowledge, there are not many other examples in the literature employing the nontilted JRES spectra to mimic a CPMG spectrum, as presented in this work.

Figure 1 shows the α ^{1}H region for the simulated equimolar mixture, with the contour plot of the ntJRES spectra for each amino acid identified by different colors, bottom plot, with the reconstructed 1D ^{1}H spectra for the individual amino acids (sum projection, with the same colors), together with their sum spectrum, top plot. A total of 14 peaks (two triplets, one quartet and one doublet of doublets) are present in the ntJRES spectrum, but due to peak overlap the p-ntJRES spectrum shows only 5 instead of those 14 peaks. This implies that 9 peaks are overlapping with other peaks. It is easy to identify a "1D" peak overlap in the ntJRES spectrum, which occurs where there are peaks with similar f2 and different f1 values. Out of the five peaks in the resulting sum spectrum only one has no clear peak overlap (3.759 ppm), the rest being combinations of 2, 3 or even 4 peaks from the ntJRES spectrum.

The performance of STOCSY over 1D spectra is limited when there is peak overlap. For (at least) two spectra presenting overlap there are two cases possible: case 1) the selected driver peak does not overlap with other peaks, and

case 2) the selected driver peak overlaps with other peaks. Figure S1A shows the (p-ntJRES)-STOCSY traces for driver peaks at 2.457 ppm (subplot A) and 3.783 ppm (subplot B), as samples of the cases described above for glutamine. Figure S1C presents the p-ntJRES for each amino acid, to allow for the identification of the different multiplets that appear correlated to the chosen driver peak. It can be easily seen that besides the multiple overlap among all four amino acids in the α ^{1}H region (as seen in Figure 1) there is also an overlap of multiplets exhibiting strong coupling from both glutamine and glutamic acid around 2.10-2.15 ppm.

When the driver peak is set at 2.475 ppm, "case 1", all the peaks on the same multiplet show high correlation values (>0.99). In contrast, the other two multiplets possess reduced correlation coefficients due to peak overlap (~ 0.7 to 0.8). The STOCSY trace with driver peak at 3.783 ppm, "case 2", presented the main correlations from glutamine peaks (~ 0.7 to 0.8), then from the alanine doublet (~0.5), slightly less from glutamic acid (~0. 45), and negligible contribution for ornithine. It is worth noting that the multiplet around 2.46 ppm presents a reduced correlation coefficient when compared to the previous case, due to the overlap at the chemical shift of the driver peak. In spite of that, a database search on a peak list from this trace will only reveal glutamine as a hit, out of the four compounds "composing" the peak, with a threshold at 0.65 (vide infra).

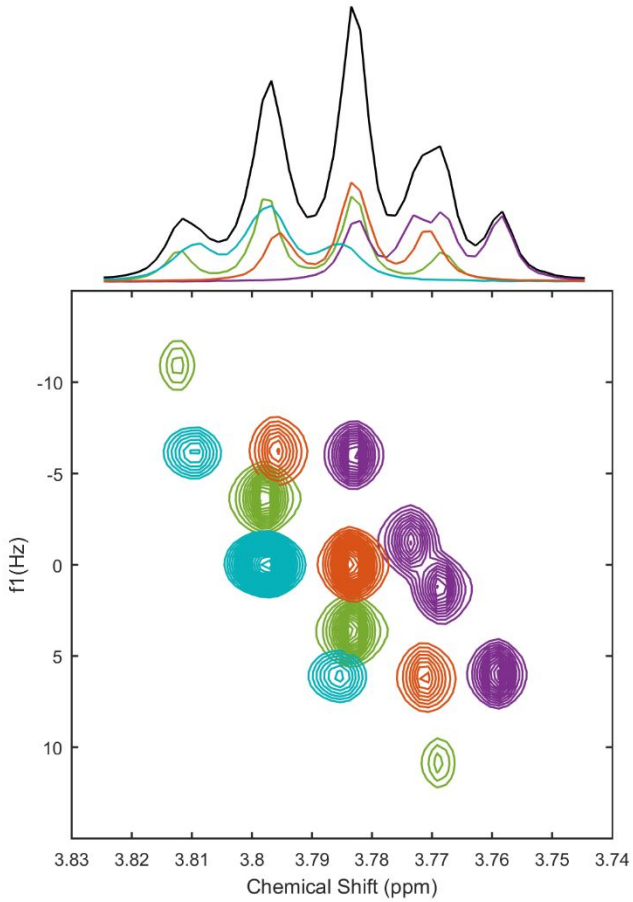


Figure 1. α ^{1}H region expansion of the ntJRES spectra from glutamine, ornithine, glutamic acid and alanine with their p-ntJRES individual spectra and sum spectrum. Bottom: ntJRES spectra from the Birmingham Metabolite Library for glutamic acid (purple), glutamine (orange), alanine (green) and ornithine (light blue). Top: p-ntJRES for each metabolite according to color in Bottom plot, and sum spectrum of the p-ntJRES in black.

An older approach to deal with peak overlap, like that of the α ^{1}H region of the simulated set, is to work with the sum projections of the 45° tilted and symmetrized JRES spectra (p-JRES) and perform a STOCSY analysis on these projections, that are similar to a homonuclear decoupled ^{1}H spectrum.[15] Figure S2 presents the α ^{1}H region for the JRES equimolar spectrum (tilted and symmetrized), with contours colored according to the different amino acids as in Figure 1. On top of this JRES spectrum the sum projection for each multiplet is plotted, now yielding a singlet, together with the overall sum of these projections.

There is no doubt that the reduction from 14 to 4 peaks in the α ^{1}H region with these projections is better for diminishing the overlap, compared to the reduction of 14 to 5 peaks in the p-ntJRES presented in Figure 1. As the centers of the multiplets are separated enough from one another and the overlap is rather small in these projections, it is expected that the application of STOCSY in these p-JRES spectra will be successful in obtaining pure STOCSY traces for each compound, ideally a trace of singlets for each.[15] However, unlike the cases of first order couplings (like these multiplets in the α ^{1}H region), it should be noted that the projections from the tilted (and symmetrized) JRES spectra of peaks with strong coupling are not clear singlets. As it will be discussed below, this has negative consequences when there is need to identify a multiplet or analyze its multiplicity to be able to extract peaks for a database search, or for structure elucidation.

Once a "projected singlet" is identified to be correlated with a chosen driver peak from another singlet, its trace should be analyzed. Figure S2 presents also the f1 traces for each one of the multiplets at their chemical shift centers. While the traces for glutamine and ornithine show clear triplets and glutamic acid presents a distorted doublet of doublets (with the two side peaks having an intensity higher than that of the two center peaks), the trace for alanine is clearly not pure. A multiplet analysis of this trace does not directly reveal a quartet, as the contribution from the two neighbor triplets is notorious and affects the trace. A critical case happens in a hypothetical overlap of multiplets where such a quartet and the triplet, as in Figure S2, overlap at the same chemical shift. Then the STOCSY analysis on the projections will suffer as a standard 1D peak overlap case. This again reinforces the need to get hold of all the information present in the 2D JRES spectra, instead of losing information while sum projecting the along f1.

Lastly, as the main objective resides in being able to generate a list of peaks for a successful database query, it is clear that working with STOCSY on the p-JRES is beneficial mostly in cases where the multiplets of the molecules have all first order couplings. The presence of multiplets with strong coupling, as evidenced on the cases of glutamic acid, glutamine and ornithine in our simulated set may have impediments in the ability of the p-JRES-STOCSY traces to create appropriate peak lists for database searches (see example for ornithine below). Furthermore, there is the additional limitation of performing a database query using only chemical shift information from the multiplet "centers", as there are no such databases available (SpinCouple requires both a chemical shift and a value for f1 in Hz). The use of STOCSY on the p-JRES spectra cannot be then thought as the optimal solution, as discussed on these examples. In spite of the success overcoming peak overlap in the mentioned cases, this methodology is not strong enough to provide appropriate peaks lists for database searches.

A substantial improvement can be obtained by analyzing the peaks in the whole ntJRES spectra, as the peaks are spread for every f2 in an added dimension, f1. Thus, the region of the amino acids α protons was analyzed not only by 5 driver peaks, from the reconstructed 1D ^{1}H spectra, or

by 4 driver peaks, from the projected "singlets", but by 14 unique driver peaks from the ntJRES that compose the four multiplets in the region. The statistical correlation analysis over the ntJRES spectra, after stretching and reshaping the data, can be presented in 2D plots in a similar way as a 1D STOCSY trace, where the color map indicates the correlation coefficients while the covariance is now shown as contour levels. A similar color-coded contour plot (aerial view with color indicating loadings) was already reported by Viant and coworkers in 2007,[31] when they performed MVDA over JRES spectra from extracts of fish liver from two rivers. An alternative display of the information on the color-coded contour of the data from Viant and coworkers is presented as a side view of the 3D loadings plot (equivalent to a backscaled loadings plot[32]). Since those spectra were tilted, the side view plot is comparable to what could be obtained by performing MVDA directly on the pJRES spectra, as reported by Verpoorte and coworkers.[33,34]

2D statistical correlation analysis of JRES data was already reported in RED-STORM and STORM (referred to as $STORM_2$), over JRES spectra (tilted and symmetrized), although the authors did not make use of the usual color map representation normally employed in 1D STOCSY. In RED-STORM the high computational load is reduced by analyzing a small set of variables, and dots at each significant correlated peak are drawn over the 2D spectrum. Besides the need for the existence of a subset of spectra without the peak overlap present, RED-STORM showed examples having only first order coupling. Although not clarified, it is evident that the retrieval of correlated peaks in first order multiplets allows for an easy construction of a peak list for database search, but this becomes a challenge when strong coupling is present. We assume a first approach would be to extract the STOCSY trace at the chemical shift corresponding to the maximum intensity in the sum projection of the JRES. Nonetheless, this trace might not be the best representation of the multiplet structure observed in a 1D experiment for the same compound, as discussed further below with the ornithine spectra as an example.

Instead, this work performs STOCSY in the ntJRES spectra, as it allows the reconstruction of a STOCSY 1D ^{1}H pseudo-spectrum mimicking a CPMG spectra of a single molecule, a trace similar to that resulting from applying STOCSY in a 1D ^{1}H spectral set. The covariance is summed over the indirect dimension and the color-coded correlation value in f2 is taken from the maximum correlation value at each f1 trace, creating the p-(ntJRES-STOCSY) trace. This allows the extraction of peaks from the 1D pseudo-spectrum by peak picking above a certain threshold, and the creation of a peak list ready for database queries in the 1D databases used when performing STOCSY on standard 1D ^{1}H spectra. The principal advantage is that the overlap that could have occurred for peaks at the same f2 is now excluded for peaks that do not correlate with the driver peak at the selected (f2, f1) pair. The methodology then compares the p-(ntJRES-STOCSY) trace with the trace obtained from the STOCSY analysis of the p-ntJRES spectra, (p-ntJRES)-STOCSY.

Figure 2 presents the COCOA-POD plot for the driver peak at f2 = 3.783 ppm and f1 = +0.008 ppm. It can be seen that the correlation reveals only peaks belonging to the alanine spectrum, the doublet around 1.47 ppm and the quartet for the α ^{1}H around 3.78 ppm. The magnified contour plot in Figure 2A reveals the 14 peaks portrayed in Figure 1 with only three other peaks, from the alanine quartet correlating with the driver peak, and the 10 other peaks presenting no correlation. It also demonstrates the presence of overlap discussed above, as the f1 trace crosses contours from one peak from each of the other three amino acids comprising the simulated set. Figure 2B reveals the projection of the STOCSY plot in Figure 2A, where the quartet and the doublet for alanine can be easily identified as the only multiplets present with high correlation, with the zoom at the driver peak region revealing the quartet structure. Finally, Figure 2C displays the STOCSY trace for the 1D STOCSY analysis over the p-ntJRES spectra with driver peak at f2 = 3.783 ppm. It is evident that the STOCSY traces from Figure 2B and Figure 2C do not coincide, the latter being identical to the one in Figure S1B. Peak picking from the (p-ntJRES)-STOCSY trace surely does not provide a search as successful as the one with the peaks from Figure 2B.

Figure S3A shows the f1 trace at f2 = 3.783 ppm corresponding to Figure 2A. The presence of several peaks in this mean ntJRES trace is an indication of peak overlap, either partial or total, in the regular 1D ^{1}H experiment. While this was shown for this f2 value where all four amino acids possess an overlapping peak each, the STOCSY analysis clearly indicates that there is no correlation to a peak other than the driver peak itself, as the STOCSY trace is almost zero in covariance and correlation for the other three peaks in the mean ntJRES f1 trace. In addition, an estimation of the degree of overlap is performed calculating the area of the correlated peak relative to the overall area under the mean ntJRES trace, in this case yielding about 25%.

Figure S3B presents the multiplet trace (45° trace) for Figure 2A. As shown in Figure S2, the alanine quartet is mixing with the triplets from ornithine and glutamine, evidenced here as well in the increase of intensity at the quartet peak valleys in the mean ntJRES. However, when the ntJRES-STOCSY trace is analyzed instead for the multiplet, the covariance and correlation in the peak valleys drops to zero and a proper multiplet analysis can be performed only on the correlated peaks, which clearly depict a quartet structure. The modified algorithm described above was employed to extract a coupling constant for the quartet of 7.4 Hz. Highlighting the correlation of the driver peak only to the peaks of the multiplet, and not to other peaks contributing to the mean intensity at the multiplet trace, allows the proper identification of the multiplet and its parameters, multiplicity and coupling constants, for first order couplings. This information is complementary to the peak list obtained and used for database query, and it needs to be contrasted with the spectrum of the proposed candidate compound to aid in structure confirmation (or elucidation).

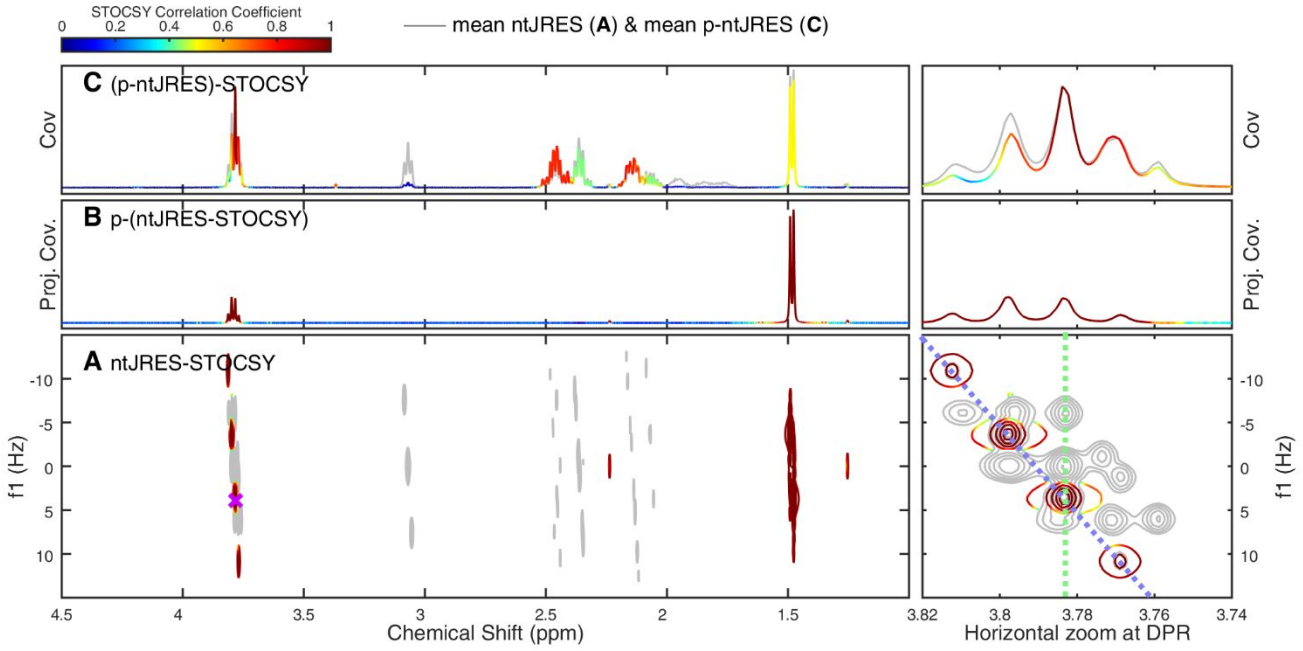


Figure 2. COCOA-POD on simulated set, f2 = 3.783 ppm, f1 = +0.008 ppm. Each subplot is provided with a Horizontal zoom at the Driver Peak Region (DPR) A) Mean ntJRES spectrum (gray) and ntJRES-STOCSY contours (color-coded according to Pearson's correlation coefficient), purple cross marks driver peak f1 and f2 coordinates; B) sum projection of ntJRES-STOCSY in A) (p-(ntJRES-STOCSY)); C) Mean p-ntJRES spectrum (gray) and (p-ntJRES)-STOCSY trace (color-coded according to Pearson's correlation coefficient).

Figure S4, Figure S5 and Figure S6 exhibit the same analysis as above for the resonances corresponding to the other 3 peaks overlapping at about 3.783 ppm, namely glutamic acid at f2 = 3.783 ppm and f1 = -0.012 ppm, glutamine at f2 = 3.783 ppm and f1 = 0.000 ppm, and ornithine at f2 = 3.783 ppm and f1 = +0.013 ppm, respectively. Table S1 includes the peaks picked for each plot from the 1D traces projected from the statistical correlation analysis on the ntJRES spectra. Queries at HMDB[4] from peaks in each list resulted in hits for alanine, glutamic acid, glutamine and ornithine with high matching ratios, as discussed below. Not only the database search can be improved considerably when compared to the application of STOCSY on 1D spectra, or even the boost that could be obtained employing POD-CAST[12] on the STOCSY data, but also the lineshape of the projections obtained with COCOA-POD after the projection from the 2D statistical correlation analysis on the ntJRES spectra are closer to the real lineshape of the 1D ^{1}H experiment. The only cases where these projected traces are not expected to resemble the "pure" spectrum from the database are: a) when the molecule spectra contains peaks that are too small and pass undetected below the baseline; b) when there is overlap within the ntJRES spectra, both in f2 and f1, either partial or full with peaks from a different molecule; and c) where there is contribution from biological correlation. Contribution from biological correlation is not an issue in simulated sets like the one employed, but appears constantly in studies with samples of biological origin, like the study on the blood serum set

below. The present approach does not offer a solution to isolate the mixed contributions to a STOCSY plot from two or more biologically (highly) correlated compounds.

Table S2 displays the database query results for the 14 peaks in the α ^{1}H region from the p-(ntJRES-STOCSY) traces compared to the results for the (p-ntJRES)-STOCSY traces, peak picking with a high threshold value (0.85) in both cases, and also with a lower threshold value (0.65) for the latter. The original p-ntJRES spectrum for each amino acid is used as a reference, as a good indication of the limitation imposed in cases of strong coupling, where the multiplet profiles are not exactly as would be in a CPMG experiment. It is worth noticing that the outcome of the searches with the p-(ntJRES-STOCSY) traces was successful in identifying the right amino acid. In contrast, for the (p-ntJRES)-STOCSY traces a lower threshold resulted either in lower JI values for the appropriate match or in wrong hits, and a high threshold value became too demanding for the query to be successful, as evidenced in most mismatches by the small number of peaks being picked.

The lower JI obtained for some molecules is not directly attributable to a failure of COCOA-POD to construct an appropriate peak list, but it is due to the procedure of construction of the peak list in the databases as well, especially for multiplets exhibiting strong coupling (it should not be a problem for spectra composed only of first order multiplets). For example, for ornithine, Figure S7 presents a superimposition for the multiplets presenting strong coupling between 1.7 and 2.0 ppm of the HMDB 1D ^{1}H spectrum, the BML 1D ^{1}H spectrum, the BML p-ntJRES spectrum and the reconstruction from the BML JRES f1 traces (selected for the maxima on the center of the three multiplets within the range shown, tilted back 90° into f2, for the p-JRES spectrum before and after symmetrization). The comparison between the ^{1}H spectra from the databases already presents slight differences in the number of peaks in each of the three multiplets at first sight. Similarly, the p-ntJRES yields a similar number of peaks than the ^{1}H spectrum from the same database, BML, and only a few differences with that from HMDB. Nonetheless, the overall profile (chemical shifts and intensities) of the p-ntJRES is a good representation of either ^{1}H spectra. In contrast, the f1 trace composites lack the intensity relations among the multiplets (2:1:1) and poorly match the chemical shifts of many peaks (obviously worse in the case of the symmetrized traces).

The peak list for ornithine lacks for example 4 peaks in the 1.85 to 1.90 ppm region, which currently fall into the unmatched peaks category. If those peaks were included it would improve the search JI values by increasing the matched and decreasing the unmatched counts. However, there are also missing peaks, as the HMDB peak list possesses a few “shoulder” peaks that passed undetected by COCOA-POD from the JRES spectral data. It might be possible to use peak picking algorithms that would detect these “shoulder” peaks, or to adjust the apodization employed on f2 to obtain more detailed projections that would have these “shoulder” peaks detectable with the current peak picking algorithm. However, the best solution

would be to match the peak picking procedure employed by the database to be queried. As this was not the scope of this work, we applied the recommended parameters by Viant and coworkers, as described above.[22] It is worth noting that, although the f1 traces from strong coupling multiplets do not reproduce appropriately the profile of the ^{1}H spectrum, a query from peaks obtained from these composite traces might in some cases yield acceptable JI values, as the success of the search depends on the the chosen tolerance and on the number of peaks within a given chemical shift range (which can be increased or diminished with the symmetrization step). It is necessary to inspect in detail the peak lists in the databases against the query lists submitted, as well as to download, process and compare the raw spectra against the STOCSY traces obtained, to advance in the identification process.[35]

COCOA-POD in blood serum set. A collection of 150 blood serum samples from volunteer donors had been analyzed by ^{1}H NMR, NOESY 1D and CPMG, and 2D ^{1}H homonuclear JRES, using the default parameters from the "Bruker Profiler" protocol. Although the initial employment of the JRES spectra is for identification purposes,[36] being experiments with only 1 scan, the mean p-ntJRES from these spectra possess far less peaks than the mean spectrum of the "Bruker Profiler" CPMG spectrum (64 scans), 89 against 251 peaks, rendering the JRES spectral set a limited ability to provide valuable information if the compounds of interest are in the low concentration range. Even the JRES with 32 scans (with an acquisition time over half an hour) had 223 peaks on the p-ntJRES, still below the number of peaks detected with the CPMG experiment for the same representative sample. Despite this limitation, the blood serum JRES spectra were analyzed with the COCOA-POD methodology, and several peaks of interest presenting peak overlap were examined in detail and discussed below.

The identification of the metabolites through database matching is only enough to reach a Level 2 of confidence, "putatively annotated compound", but to properly identify a compound, Level 1, the use of at least two independent and orthogonal analytical techniques is necessary, applied both to the metabolite of interest and to an authentic reference standard.[29] For example, in NMR, this data from the database can be complemented with the information in the JRES spectra acquired, as the information for the multiplet coupling constants must match those in databases as well. Without performing any experiment with reference standards, we were able to annotate peaks (putative identification), more precisely whole ntJRES-STOCSY (multiplet structures) and p-(ntJRES-STOCSY) traces to almost 20 of the most abundant compounds found in blood serum, as listed in Table S3, ordered by their mean concentration, as reported elsewhere.[37]

Additional two-dimensional spectra can be obtained to complete the process of structure elucidation or confirmation (in most cases). Experiments can be either homonuclear, like COSY and TOCSY, or heteronuclear, like HSQC and HMBC, to name the ones employed more frequently.[35] It might be useful and/or necessary to expand the analysis employing a different technique, for example

mass spectrometry for obtaining an accurate molecular mass and molecular formula. The use of the authentic reference standard in spiking experiments also aids in the confirmation of the identity, as the signals expected to belong to the specific compound should increase after the addition of a standard (at an appropriate concentration for the increase to be observed, but not in excess as to obscure the original signals).[35]

The number of potential peak overlaps is approximated by the difference between the total amount of peaks distinguished in the JRES spectra and the number of peaks detected in the p-ntJRES. In this blood serum representative sample with only 1 scan this number was really low, with only 11 overlapping peaks detected. In the blood serum spectra an interesting case of peak overlap is that of alanine, as displayed in Figure 3, with the driver peak being the most upfield doublet peak at f2 = 1.471 ppm and f1 = +0.006 ppm. The contour plot (Figure 3A) shows clear high correlation to the other peak from the doublet, together with the quartet peaks at 3.78 ppm. These correlations are of course translated into the projection in Figure 3B, but the quartet is absent in the 1D STOCSY analysis over the projections in Figure 3C. This is an indication that some peaks are being lost with negligible correlation beneath the glucose peaks in the 3.78 ppm region. The choice of the other doublet peak from alanine provides a similar result.

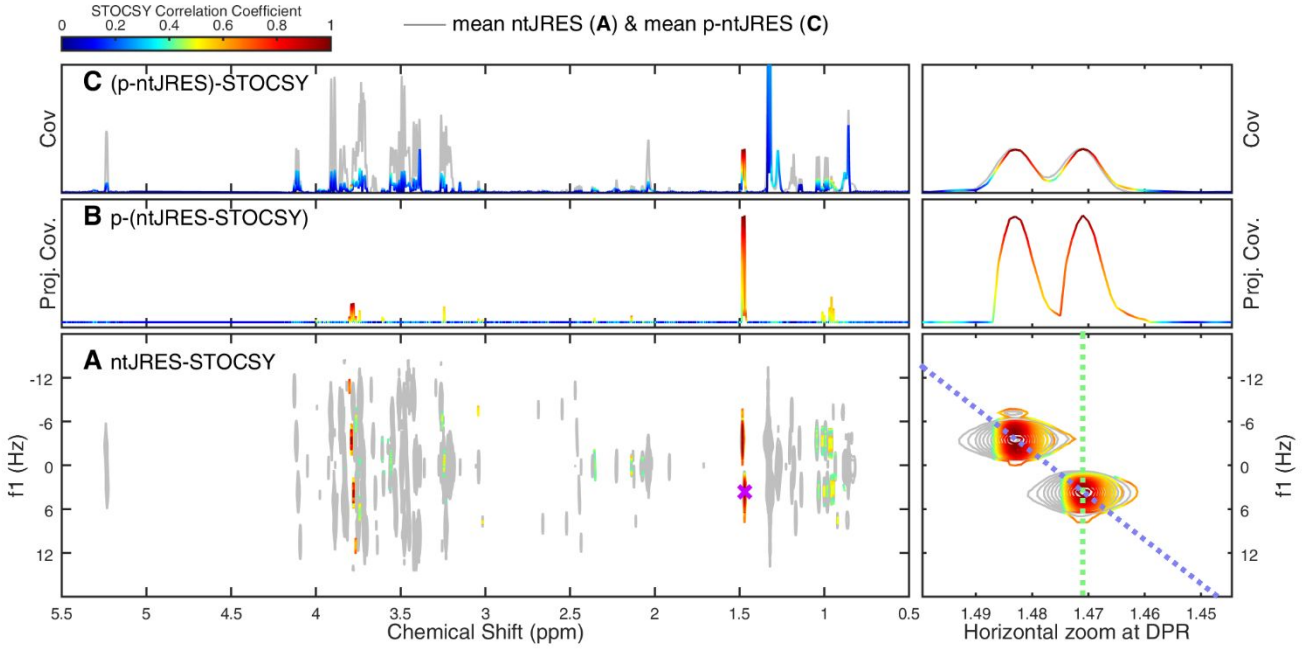


Figure 3. COCOA-POD on blood serum set, f2 = 1.471 ppm, f1 = +0.006 ppm. Each subplot is provided with a Horizontal zoom at the Driver Peak Region (DPR) A) Mean ntJRES spectrum (gray) and ntJRES-STOCSY contours (color-coded according to Pearson's correlation coefficient), purple cross marks driver peak f1 and f2 coordinates; B) sum projection of ntJRES-STOCSY in A) (p-(ntJRES-STOCSY)); C) Mean p-ntJRES spectrum (gray) and (p-ntJRES)-STOCSY trace (color-coded according to Pearson's correlation coefficient).

Reversely, if one of the peaks from the alanine quartet is chosen as the driver peak, for example that at f2 = 3.778 ppm and f1 = +0.006 ppm, as shown in Figure S8, it can be clearly observed that the correlations from the 2D STOCSY

analysis and its projection remain similar to those in Figure 3, while the 1D STOCSY analysis over the projection resembles that of any trace with the glucose peak being chosen as driver peak. It is worth noticing the high degree of overlap in Figure S8A (horizontal zoom at DPR) and the absence of correlation to the glucose peaks at the same f2 value. The query on HMDB with the 6 peaks obtained in Figure S8B gave alanine as a first hit with JI equal to 1. This result would have never been attainable with a 1D STOCSY analysis or even with the aid of POD-CAST applied to that 1D STOCSY analysis,[12] as this case has contribution to the correlation from the glucose peaks only, and not a mixed contribution from peaks corresponding to both metabolites (which is essential for POD-CAST to account for the overlap).

Another peak overlap occurs at f2 = 1.192 ppm, where a side peak from the ethanol triplet overlaps with one peak from the 3-hydroxybutyrate doublet. Figure 4 shows the COCOA-POD for both driver peaks, ethanol at f2 = 1.193 ppm and f1 = -0.012 ppm, top, and 3-hydroxybutyrate at f2 = 1.191 ppm and f1 = +0.006 ppm, bottom. The analysis for ethanol is simple, the contour of the 2D STOCSY (Figure 4A'), its projection (Figure 4B') and the 1D STOCSY (Figure 4C) all have the ethanol peaks. It is hard to realize that a peak with low contribution to correlation on the 1D STOCSY overlaps there at 1.180 ppm, unless the magnification of the driver peak region is inspected on the ntJRES contour plots, where two peaks are clearly identified on the 1.180 ppm f1 trace (green dotted lines). When analyzing Figure 4A, the contour of the 2D STOCSY and its projection in Figure 4B show the peaks for 3-hydroxybutyrate, the doublet containing the driver peak, the two doublet of doublets around 2.36 ppm and even for the low intensity peaks around 4.1 ppm (although it is an incomplete multiplet that should present 6 peaks instead of 4). Also, there is correlation to the singlets for acetone and acetoacetate (due to biological correlation[38]). In contrast, the 1D STOCSY on the projections from ntJRES spectra in Figure 4C reveals the profile of the ethanol spectrum, indicating that its correlation surpasses that of the 3-hydroxybutyrate. It is evident that learning the identity of an unknown molecule would be more difficult if peaks are missing when building the peak list for database search or structural elucidation, like in the previous analyses for alanine and 3-hydroxybutyrate.

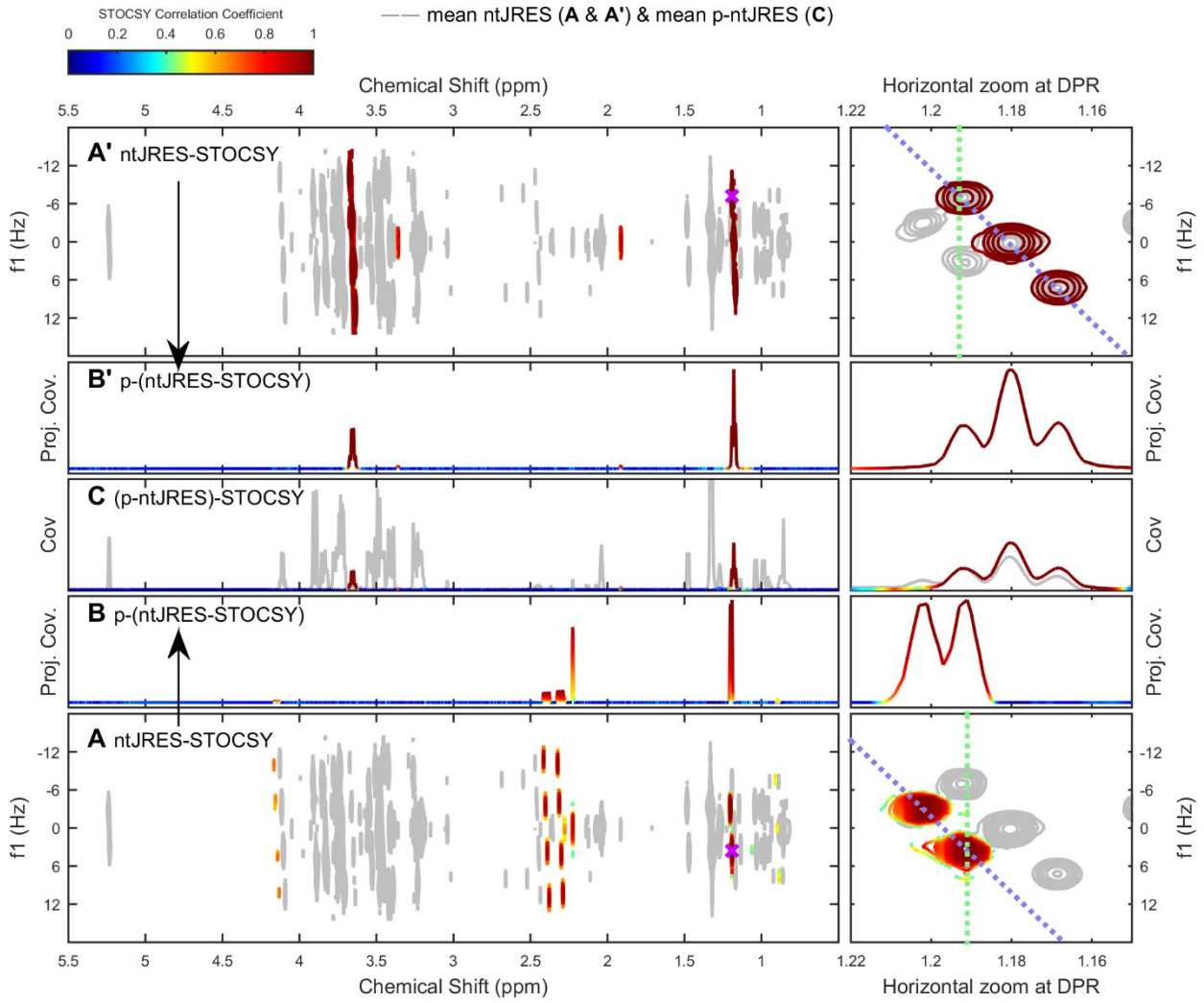


Figure 4. COCOA-POD on blood serum set for peaks around f2 = 1.19 ppm. Each subplot is provided with a Horizontal zoom at the Driver Peak Region (DPR). A and A') Mean ntJRES spectrum (gray) and ntJRES-STOCSY contours (color-coded according to Pearson's correlation coefficient), purple cross marks driver peak f1 and f2 coordinates; B and B') sum projection of ntJRES-STOCSY in A and A', respectively (p-(ntJRES-STOCSY)); C) Mean p-ntJRES spectrum (gray) and (p-ntJRES)-STOCSY trace (color-coded according to Pearson's correlation coefficient).

In addition, there is another advantage when analyzing the multiplets in the p-(ntJRES-STOCSY) pseudo-spectrum or 45° traces, against the (symmetrized) f1 traces of p-JRES (as in JRES-STOCSY or RED-STORM). If there are nuclei exhibiting first order coupling having chemical shift differences not much larger than their coupling constant, a slight distortion of the peak intensities due to the leaning effect is normally seen. The symmetrization step will remove that intensity distortion, which is useful information for the structure elucidation step in cases where the molecules are not found in databases. Such is the case of the two doublet of doublets mentioned for 3-hydroxybutyrate, around 2.36 ppm, which indicates those nuclei are coupled.

It should not be expected for peak overlap to be a problem only between molecules, as peak overlap by peaks from the same molecule also occurs. The most obvious case is that where complex multiplets overlap, but it can also happen with simple first order multiplets. One such example, very common in metabolomics, is the molecule

of glucose, with its added equilibrium between the α and β forms. Another example is that of the fatty acids, where long chains of methylenes have their resonances at around the same chemical shift. An example worth mentioning is that of the diastereotopic methyl groups of leucine, whose signals in the 1D ^{1}H spectrum can be easily mistaken for a "triplet". The two doublets are close in chemical shift, at a distance similar to the *J* coupling value of about 6.1 Hz, resembling the lineshape of a triplet with the appropriate 1:2:1 integration ratio when sum projected from the ntJRES-STOCSY spectrum. A mistaken triplet instead of two doublets clearly leads to a faulty structure elucidation process, as the leucine structure is not consistent with a methyl group linked to a methylene, but to two methyl groups linked to a methine.

A key example of the need to select the peaks within the ntJRES spectrum is the COCOA-POD plot for betaine, a molecule with only two singlets overlapping with more intense peaks from glucose, as seen in Figure S9. While the selection of the driver peaks can be oriented by the result of a MVDA on the 1D spectra set, as it was presented in RED-STORM,[18] the idea of detecting an effect from a molecule whose only two peaks are buried under peaks from the much concentrated glucose molecule seems challenging. The identification of the betaine molecule in the blood serum sample can only be performed if a peak picking routine is applied spectra wide, unless the choice of any of the two peaks as a driver peak is driven by MVDA on the 2D JRES data, like Viant and coworkers.[31] We suggest though that the MVDA be performed on the ntJRES, as its loadings (or coefficients) projection can be compared more easily with 1D ^{1}H spectral data than the projected loadings from the tJRES,[31] being always able to tilt the JRES spectrum for the specific multiplet analysis. We thus envision in the future the implementation of full 2D JRES data being analyzed by MVDA as an additional metabolomics tool, despite the quantitation issue on this type of experiment.

It is more than evident that the extent at which the JRES can help with the identification of the peaks in the spectrum relies on the number of detectable peaks. There is no doubt that 1 scan is not enough for this purpose, as for example there are no peaks for this blood serum set in the mean JRES spectra down field from the α glucose anomeric ^{1}H peak at 5.233 ppm. Figure S10 shows the number of peaks detected in the sum projection of the ntJRES spectra for experiments with increasing number of scans (1, 2, 4, 8, 16 and 32 scans) for a representative sample in the set, as well as the number of overlapping peaks, together with the acquisition time. As expected, the number of peaks in the projections increased with more scans and also the number of peak overlaps, as the small peaks begin differentiating from the noise with more scans in the experiments and overlap with higher intensity peaks already detected in the low scans cases. While several studies were reported using JRES spectra for purposes like identification, quantification or profiling, there is no consensus to the number of scans that are appropriate to acquire. The decision is of course a compromise between how much time could be added per sample to the regular

analysis and how many peaks out of the total in the CPMG or 1D NOESY are accounted for in the JRES spectra. A representative sample of a study can be evaluated at increasing number of scans to decide on an acceptable compromise before acquiring the complete set.

CONCLUSION

The spectra involved in metabolomics studies normally come from complex mixtures and present hundreds of peaks, with a high probability of peak overlap within them. STOCSY is one of the main tools to help with the identification of compounds in ^{1}H NMR-based metabolic phenotyping studies, being able to create a peak list as input for the database query based on correlation of peaks to a chosen driver peak. A simple and quick methodology was devised to account for and overcome the peak overlap issue in ^{1}H NMR metabolomics studies when employing STOCSY. The statistical correlation analysis was performed over the ntJRES spectra set, taking advantage of the spread of the peaks in the indirect dimension, as recently introduced in RED-STORM.[18] In contrast, we proposed the use of STOCSY on the whole set, without selecting a subset based on reference matching. The utilization of the spectra without the standard 45° tilt processing step allowed us to create a sum projection of the 2D STOCSY covariance-correlation matrix to obtain a 1D STOCSY trace, p-(ntJRES-STOCSY), which in cases of overlap at the f2 value of the driver peak in a regular 1D experiment is lacking the contribution of correlation from other peaks at the same f2 value. Overlap contribution is thus limited only to cases of overlap in both f2 and f1, which occurs far less frequently, and can be assessed with other approaches like POD-CAST.[12]

The COCOA-POD methodology compares the f1 trace from the mean ntJRES spectrum at the driver peak f2 value with the f1 trace from the ntJRES-STOCSY at the same f2 value, to account for overlap with peaks from other compound (or even the same). Analogously, the 45° trace for the multiplet at the mean ntJRES is compared with the 45° trace from the ntJRES-STOCSY to evaluate the multiplet contamination due to neighboring multiplets. The correlation trace for the multiplet is analyzed with an algorithm to determine its multiplicity and coupling constants, in cases of first order multiplets.[28]

The p-(ntJRES-STOCSY) traces obtained for each driver peak can be subjected to a peak picking routine to create a peak list for database query. This peak list would certainly outperform the peak list originated from a 1D STOCSY analysis in a database search comparison, as the latter would clearly suffer from the peak overlap consequences (known for long when employing STOCSY) presenting masked or mixed correlations, yielding either incomplete or erroneous peak lists in the first case, or expanded lists in the latter. Having obtained an unsuccessful search, the process for identification should proceed with more emphasis on the additional 2D experiments available, to pursue either the structure elucidation of an unknown compound or the identification confirmation of a known molecule.[35]

The method developed, COCOA-POD, outcomes the use of STOCSY on projections from JRES spectra (p-JRES)-STOCSY as a way to deal with peak overlap. The latter is limited by multiplets that are too close to each other, as well as it exhibits difficulties in interpretation when having multiplets whose coupling is not of first order. This is also a complication when a strong coupling multiplet is identified by RED-STORM, even more if a database query is to be performed with its outcome. In contrast, the application of STOCSY on nontilted projections in COCOA-POD concludes with a searchable peak list. If the molecule is unknown, either COCOA-POD, (p-JRES)-STOCSY and RED-STORM would indicate the multiplet is of high order, being able to use that information for structure elucidation.

Complex mixtures as such are not exclusive to metabolomics studies. In fact, the utilization of COCOA-POD is suggested for any set, whether the matrix has been broadly described or not in the past, as the generated traces can be used to perform structural elucidation (assuming the correlated peaks are from a single molecule) even if there is no reference spectrum from a database to compare against. It is expected that the use of JRES is expanded in the future of metabolomics studies, not only as a tool for identification purposes, but also for MVDA on the set. The employment of nontilted JRES spectra is suggested also for the MVDA step, being able to obtain projected loadings that resemble 1D ^{1}H spectra, instead of the reported tilted projected loadings.[33,34]

ASSOCIATED CONTENT

Supporting Information. The Supporting Information is available free of charge on the ACS Publication Website at DOI:

Section 1: Multiple testing correction. Scheme S1. Procedure for generating COCOA-POD subplots A, B and C. Pseudocode for the algorithm. Section 2: Figure S1. (p-ntJRES)-STOCSY of simulated data set with driver peaks at 2.457 and 3.783 ppm, and p-ntJRES spectra for each amino acid. Figure S2. JRES of the α ^{1}H region of the amino acids, p-JRES spectra for the amino acids and sum spectrum, and f1 traces for the α ^{1}H multiplet centers from JRES. Figure S3. Analysis of traces for COCOA-POD on simulated spectra set, f2 = 3.783 ppm, f1 = +0.008 ppm. Figure S4. COCOA-POD on simulated spectra set, f2 = 3.783 ppm, f1 = -0.012 ppm. Figure S5. COCOA-POD on simulated spectra set, f2 = 3.783 ppm, f1 = 0.000 ppm. Figure S6. COCOA-POD on simulated spectra set, f2 = 3.783 ppm, f1 = +0.013 ppm. Table S1. Chemical shifts for all peaks found in p-(ntJRES-STOCSY) traces from the four overlapping peaks at f2 = 3.783 ppm in the simulated spectra set. Table S2. Database search in HMDB for the simulated spectra set for the 14 driver peaks in the α ^{1}H region of the amino acids. Figure S7. Trace comparison for ornithine multiplets between 1.70 and 2.00 ppm from ^{1}H 1D experiments, from f2 projection of p-ntJRES and from f1 traces of JRES. Table S3. List of annotated compounds for the blood serum set with a match in the database query. Figure S8. COCOA-POD on blood serum set, f2 = 3.778 ppm, f1 = +0.006 ppm. Figure S9. COCOA-POD on blood serum set, f2 = 3.260 ppm, f1 = 0.000 ppm. Figure S10. Number of total peaks in p-ntJRES, number of overlapping

peaks on ntJRES and experiment time with increasing number of scans.

AUTHOR INFORMATION

Corresponding Author

* E-mail: pablo.hoijemberg@cibion.conicet.gov.ar.

Notes

The authors declare no competing financial interest.

Author Contributions

⊥A.C.M. and G.R. contributed equally to this work.

ACKNOWLEDGMENTS

The authors thank ANPCYT (through project # PICT-PRH 2016-0014) and ASaCTeI (Santa Fe, Argentina) for the financial support.

A.C.-M. and G.R. thank CONICET for their scholarships (Becas doctorales internas Temas Estratégicos).

The authors thank CIBIC Lab for sample collection and Andrea V. Coscia for spectral data acquisition.

REFERENCES

(1) Lindon, J. C.; Holmes, E.; Nicholson, J. K. System Biology: Metabolomics. FEBS J. 2007, 274 (5), 1140–1151.

(2) Nicholson, J. K.; Lindon, J. C. Systems biology: Metabonomics. Nature 2008, 455 (7216), 1054–1056.

(3) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; et al. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets. Anal. Chem. 2005, 77 (5), 1282–1289.

(4) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. HMDB 3.0-The Human Metabolome Database in 2013. Nucleic Acids Res. 2013, 41 (D1), 801–807.

(5) 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 (suppl_1), D402–D408.

(6) Robinette, S. L.; Lindon, J. C.; Nicholson, J. K. Statistical spectroscopic tools for biomarker discovery and systems medicine. Anal. Chem. 2013, 85 (11), 5297–5303.

(7) Wei, S. W.; Zhang, J.; Liu, L. Y.; Ye, T.; Gowda, G. a N.; Tayyari, F.; Raftery, D. Ratio Analysis Nuclear Magnetic Resonance Spectroscopy for Selective Metabolite Identification in Complex Samples. Anal. Chem. 2011, 83 (20), 7616–7623.

(8) Vu, T. N.; Laukens, K. Getting your peaks in line: A review of alignment methods for NMR spectral data. Metabolites 2013, 3 (2), 259–276.

(9) Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nat Protoc 2007, 2 (11), 2692–2703.

(10) Yuan, J.; Zhang, B.; Wang, C.; Bruschweiler, R. Carbohydrate background removal in metabolomics samples. Anal. Chem. 2018, 90 (24), 14100–14104.

(11) Posma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M. D.; Nicholson, J. K. Subset optimization by reference matching (STORM): An optimized statistical approach for recovery of metabolic biomarker structural information from ^{1}H NMR spectra of biofluids. Anal. Chem. 2012, 84 (24), 10694–10701.

(12) Hoijemberg, P. A.; Pelczer, I. Fast Metabolite Identification in Nuclear Magnetic Resonance Metabolomic Studies: Statistical Peak Sorting and Peak Overlap Detection for More Reliable Database Queries. J. Proteome Res. 2018, 17 (1), 392–401.

(13) Öman, T.; Tessem, M.; Bathen, T. F.; Bertilsson, H.; Angelsen, A.; Hedenström, M.; Andreassen, T. Identification of

metabolites from 2D ^{1}H- ^{13}C HSQC NMR using peak correlation plots. 2014, 1–8.

(14) Rudd, T. R.; Macchi, E.; Muzi, L.; Ferro, M.; Gaudesi, D.; Torri, G.; Casu, B.; Guerrini, M.; Yates, E. A. Unravelling structural information from complex mixtures utilizing correlation spectroscopy applied to HSQC spectra. Anal. Chem. 2013, 85 (15), 7487–7493.

(15) Johnson, C. H.; Athersuch, T. J.; Wilson, I. D.; Iddon, L.; Meng, X.; Stachulski, A. V.; Lindon, J. C.; Nicholson, J. K. Kinetic and *J*-resolved statistical total correlation NMR spectroscopy approaches to structural information recovery in complex reacting mixtures: Application to acyl glucuronide intramolecular transacylation reactions. Anal. Chem. 2008, 80 (13), 4886–4895.

(16) Zangger, K. Pure shift NMR. Prog. Nucl. Magn. Reson. Spectrosc. 2015, 86–87, 1–20.

(17) Sinnaeve, D.; Foroozandeh, M.; Nilsson, M.; Morris, G. A. A general method for extracting individual coupling constants from crowded ^{1}H NMR spectra. Angew. Chemie - Int. Ed. 2016, 55 (3), 1090–1093.

(18) Posma, J. M.; Garcia-Perez, I.; Heaton, J. C.; Burdisso, P.; Mathers, J. C.; Draper, J.; Lewis, M.; Lindon, J. C.; Frost, G.; Holmes, E.; et al. Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. Anal. Chem. 2017, 89 (6), 3300–3309.

(19) 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 (1), 659–665.

(20) 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 ^{1}H and 2-D ^{1}H *J*-resolved NMR spectra of authentic metabolite standards (BML-NMR). Metabolomics 2012, 8 (1), 8–18.

(21) Robinette, S. L.; Zhang, F.; Brüschweiler-Li, L.; Brüschweiler, R. Web server based complex mixture analysis by NMR. Anal. Chem. 2008, 80 (10), 3606–3611.

(22) Parsons, H. M.; Ludwig, C.; Viant, M. R. Line-shape analysis of *J*-resolved NMR spectra: application to metabolomics and quantification of intensity errors from signal processing and high signal congestion. Magn. Reson. Chem. 2009, 47 (S1), S86--S95.

(23) Meiboom, S.; Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. Rev. Sci. Instrum. 1958, 29, 688–691.

(24) Carr, H. Y.; Purcell, E. M. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. Phys. Rev. 1954, 94 (3), 630–638.

(25) Nicholson, J. K.; Foxall, P. J. D.; Spraul, M.; Farrant, R. D.; Lindon, J. C. 750-Mhz H-1 and H-1-C-13 NMR-Spectroscopy of Human Blood-Plasma. Anal. Chem. 1995, 67 (2), 793–811.

(26) Alves, A. C.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. Analytic properties of statistical total correlation spectroscopy based information recovery in ^{1}H NMR metabolic data sets. Anal. Chem. 2009, 81 (6), 2075–2084.

(27) Newbold, P.; Carlson, W. L.; Thorne, B. M. Statistics for business and economics, 8th ed.; Pearson Education: Upper Saddle River, NJ, 2013 ; pp 432-434.

(28) Hoye, T. R.; Zhao, H. A Method for Easily Determining Coupling Constant Values: An Addendum to "A Practical Guide to First-Order Multiplet Analysis in ^{1}H NMR Spectroscopy." J. Org. Chem. 2002, 67 (12), 4014–4016.

(29) Viant, M. R.; Kurland, I. J.; Jones, M. R.; Dunn, W. B. How close are we to complete annotation of metabolomes? Curr. Opin. Chem. Biol. 2017, 36, 64–69.

(30) Shapiro, M. J.; Chin, J.; Marti, R. E.; Jarosinski, M. A. Enhanced resolution in MAS NMR for combinatorial chemistry. Tetrahedron Lett. 1997, 38 (8), 1333–1336.

(31) Parsons, H. M.; Ludwig, C.; Günther, U. L.; Viant, M. R. Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. BMC Bioinformatics 2007, 8 (1), 234.

(32) Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies. Anal. Chem. 2005, 77 (2), 517–526.

(33) Liang, Y.; Hae, Y.; Kim, H. K.; Linthorst, H. J. M.; Verpoorte, R. Metabolomic analysis of methyl jasmonate treated Brassica rapa leaves by 2-dimensional NMR spectroscopy. Phytochemistry 2006, 67, 2503–2511.

(34) Choi, Y. H.; Kim, H. K.; Linthorst, H. J. M.; Hollander, J. G.; Lefeber, A. W. M.; Erkelens, C.; Nuzillard, J.; Verpoorte, R. NMR Metabolomics to Revisit the Tobacco Mosaic Virus Infection in Nicotiana tabacum Leaves. 2006, 742–748.

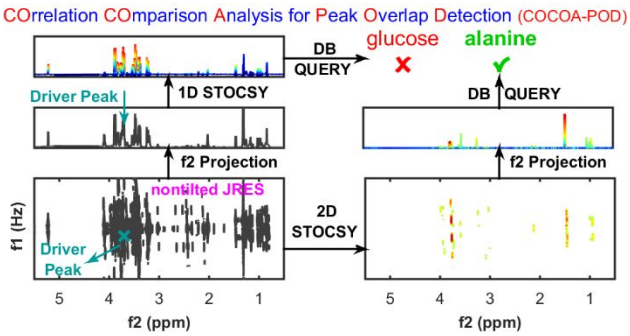
(35) Dona, A. C.; Kyriakides, M.; Scott, F.; Shephard, E. A.; Varshavi, D.; Veselkov, K.; Everett, J. R. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput. Struct. Biotechnol. J. 2016, 14, 135–153.

(36) Ludwig, C.; Viant, M. R. Two-dimensional *J*-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. Phytochem. Anal. 2010, 21 (1), 22–32.

(37) Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; et al. The human serum metabolome. PLoS One 2011, 6 (2), e16957.

(38) 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, 89 (21), 11405–11412.

for TOC only


COrrelation COmparison Analysis for Peak Overlap Detection (COCOA-POD)
DB QUERY
glucose
alanine
1D STOCSY
Driver Peak
DB QUERY
f2 Projection
f2 Projection
nontilted JRES
f1 (Hz)
Driver Peak
2D STOCSY
5 4 3 2 1
f2 (ppm)
5 4 3 2 1
f2 (ppm)

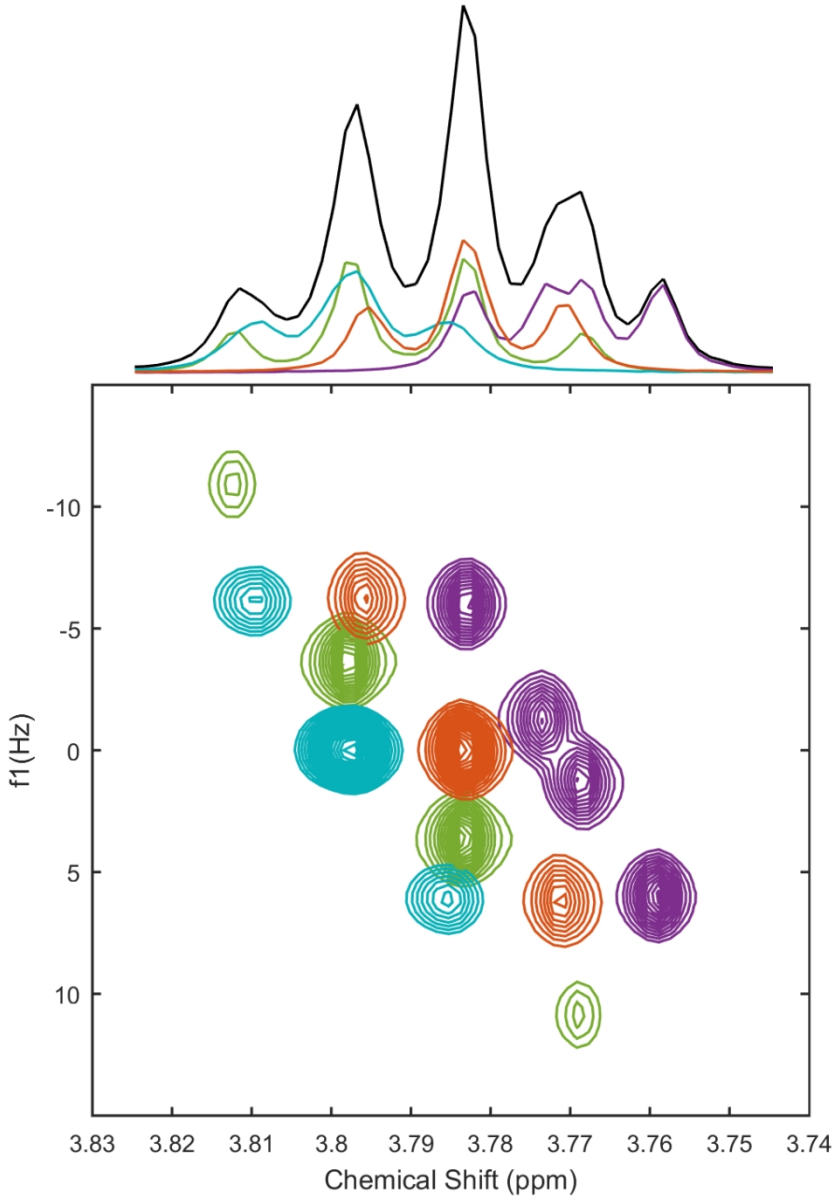


Figure 1. α ^{1}H region expansion of the ntJRES spectra from glutamine, ornithine, glutamic acid and alanine with their p-ntJRES individual spectra and sum spectrum. Bottom: ntJRES spectra from the Birmingham Metabolite Library for glutamic acid (purple), glutamine (orange), alanine (green) and ornithine (light blue). Top: p-ntJRES for each metabolite according to color in Bottom plot, and sum spectrum of the p-ntJRES in black.

87x123mm (300 x 300 DPI)

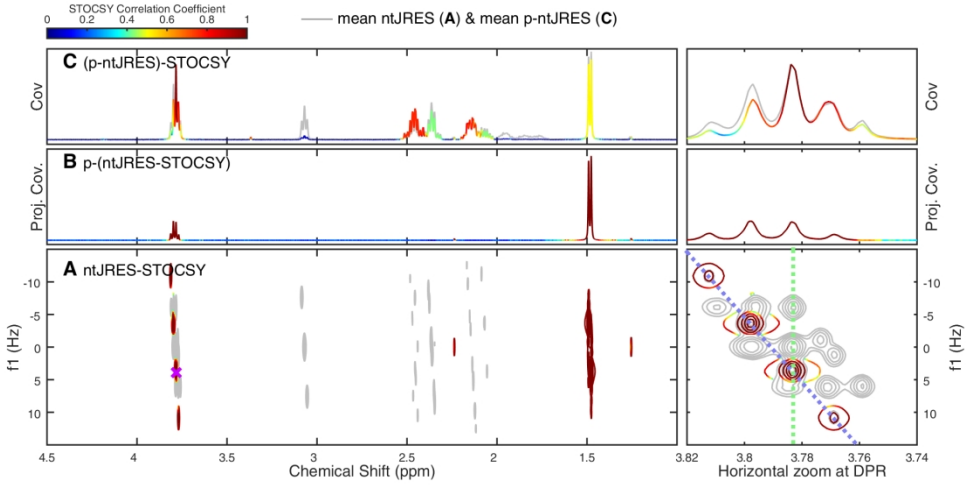


Figure 2. COCOA-POD on simulated set, f2 = 3.783 ppm, f1 = +0.008 ppm. Each subplot is provided with a Horizontal zoom at the Driver Peak Region (DPR) A) Mean ntJRES spectrum (gray) and ntJRES-STOCSY contours (color-coded according to Pear-son's correlation coefficient), purple cross marks driver peak f1 and f2 coordinates; B) sum projection of ntJRES-STOCSY in A) (p-(ntJRES-STOCSY)); C) Mean p-ntJRES spectrum (gray) and (p-ntJRES)-STOCSY trace (color-coded according to Pearson's correlation coefficient).

177x88mm (300 x 300 DPI)

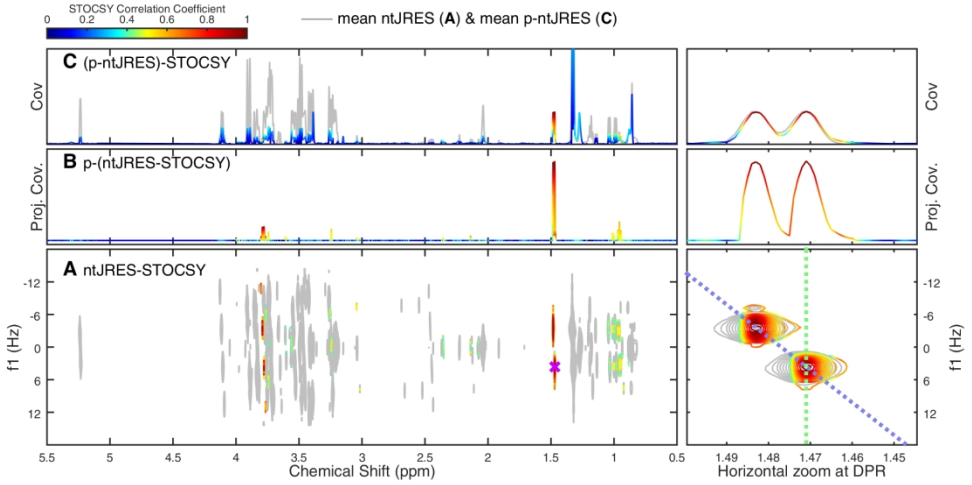


Figure 3. COCOA-POD on blood serum set, f2 = 1.471 ppm, f1 = +0.006 ppm. Each subplot is provided with a Horizontal zoom at the Driver Peak Region (DPR) A) Mean ntJRES spectrum (gray) and ntJRES-STOCSY contours (color-coded according to Pearson's correlation coefficient), purple cross marks driver peak f1 and f2 coordinates; B) sum projection of ntJRES-STOCSY in A) (p-(ntJRES-STOCSY)); C) Mean p-ntJRES spectrum (gray) and (p-ntJRES)-STOCSY trace (color-coded according to Pear-son's correlation coefficient).

177x88mm (300 x 300 DPI)

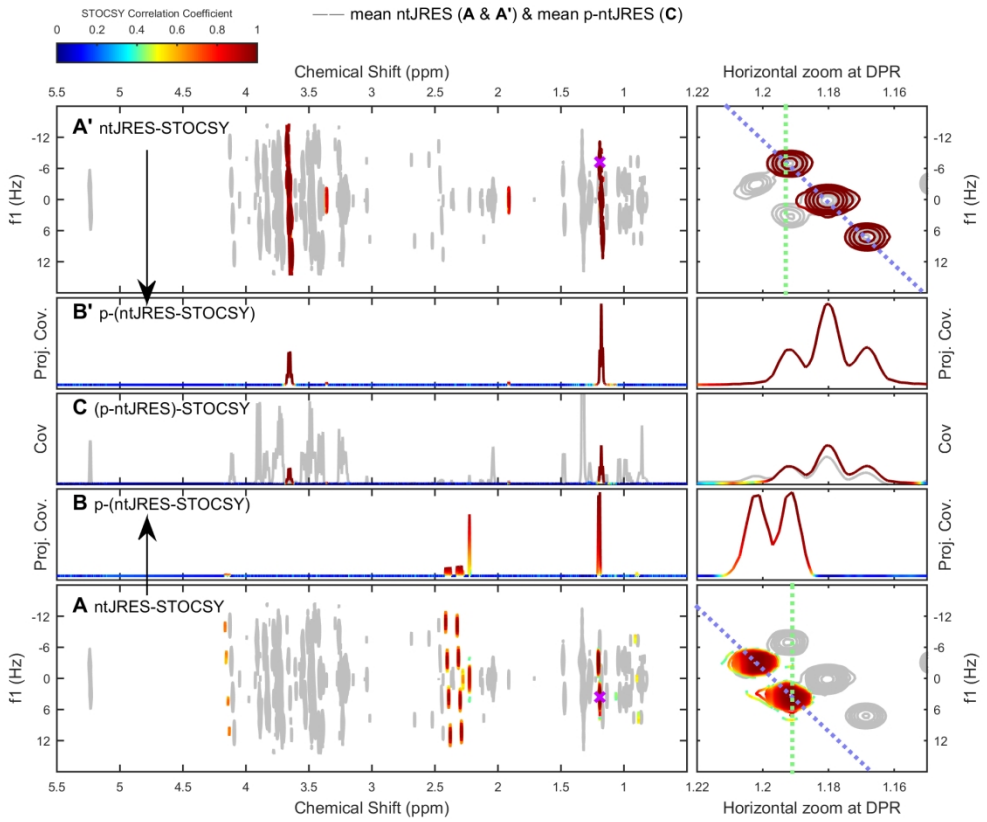


Figure 4. COCOA-POD on blood serum set for peaks around f2 = 1.19 ppm. Each subplot is provided with a Horizontal zoom at the Driver Peak Region (DPR). A and A') Mean ntJRES spectrum (gray) and ntJRES-STOCSY contours (color-coded accord-ing to Pearson's correlation coefficient), purple cross marks driver peak f1 and f2 coordinates; B and B') sum projection of ntJRES-STOCSY in A and A', respectively (p-(ntJRES-STOCSY)); C) Mean p-ntJRES spectrum (gray) and (p-ntJRES)-STOCSY trace (color-coded according to Pearson's correlation coefficient).

168x137mm (300 x 300 DPI)