Tutorial: guidelines for identifying unknown metabolites using NMR-based metabolic profiling techniques

Isabel Garcia-Perez[1], Joram M. Posma[2,3], Jose Ivan Serrano-Contreras[1], Claire L. Boulangé[1], Queenie Chan[4], Gary Frost[1], Jeremiah Stamler[5], Paul Elliott[3,4,6,7], John C. Lindon[1], Elaine Holmes[1,7,8*], Jeremy K Nicholson[8*]

[1] Division of Digestive Diseases, Department of Metabolism, Digestion and Reproduction, Faculty of Medicine, Hammersmith Campus, Imperial College London, W12 0NN, U.K.;
[2] Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Faculty of Medicine, South Kensington Campus, Imperial College London, SW7 2AZ, U.K.;
[3] Health Data Research UK-London, U.K.;
[4] Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, St. Mary's Campus, Imperial College London, W2 1PG, U.K.;
[5] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, U.S.A.;
[6] MRC Centre for Environment and Health, School of Public Health, Faculty of Medicine, St. Mary's Campus, Imperial College London, W2 1PG, U.K.;
[7] Dementia Research Institute at Imperial College, Faculty of Medicine, Hammersmith Campus, Imperial College London, W12 0NN, U.K.;
[8] The Australian National Phenome Center, Harry Perkins Institute, Murdoch University, WA 6150, Australia.

***Authors for correspondence**
Elaine Holmes, elaine.holmes@imperial.ac.uk;
Jeremy K. Nicholson, jeremy.nicholson@murdoch.edu.au

ABSTRACT

Metabolic profiling of biological samples provides important insights into multiple physiological and pathological processes, but is hindered by a lack of automated annotation and standardised methods for structure elucidation of candidate disease biomarkers. Here, we describe a system for identifying molecular species derived from NMR spectroscopy-based metabolic phenotyping studies, with detailed info on sample preparation, data acquisition, and data modelling. We provide eight different modular workflows to be followed

in a recommended sequential order according to their level of difficulty. This multi-platform system involves the use of statistical spectroscopic tools such as STOCSY, STORM and RED-STORM to identify other signals in the NMR spectra relating to the same molecule. It also utilizes 2D-NMR spectroscopic analysis, separation and pre-concentration techniques, multiple hyphenated analytical platforms and data extraction from existing databases. The complete system, using all eight workflows, would take up to a month, as it includes multidimensional NMR experiments that require prolonged experiment times. However, easier identification cases using fewer steps would take two or three days. This approach to biomarker discovery is efficient, cost-effective and offers increased chemical space coverage of the metabolome, resulting in faster and more accurate assignment of NMR-generated biomarkers arising from metabolic phenotyping studies. Finally, it requires basic understanding of Matlab in order to perform statistical spectroscopic tools and analytical skills to perform Solid Phase Extraction, LC-fraction collection, LC-NMR-MS and 1D and 2D NMR experiments.

INTRODUCTION

Nuclear Magnetic Resonance (NMR) spectroscopy has been widely applied in metabolic profiling and phenotyping[1-3] for over three decades. The technology allows for accurate high-throughput screening of thousands of metabolites (small molecular species <1 kDa) present in a biological sample[4-7] such as urine, plasma, faeces, saliva and multiple types of tissues, as well as food[8] and plant extracts. NMR spectroscopy provides robust multi-metabolite fingerprints of hundreds of metabolites in many biofluids, many of which are listed in spectral databases, particularly for common biofluids in urine and blood.

However, it is very challenging to elucidate the structure of all metabolites present in biofluid samples. The large number of unknown or unidentified metabolites with high dynamic concentration range, extensive chemical diversity and different physical properties poses a substantial analytical challenge. Metabolic profiling studies are often geared towards finding differences in the levels of metabolites that are statistically correlated with a clinical outcome, dietary intervention or toxic exposure when compared with a control group. The chemical assignment of this reduced panel of biologically-relevant metabolites is possible using statistical spectroscopic tools[9-11], 2D-NMR spectroscopic analysis[12-14], separation and pre-concentration techniques[11], various chromatographic and MS-based analytical platforms[15,16] and existing spectral databases. However, the structural elucidation of NMR resonances relating to unknown molecules remains a major bottleneck in metabolic profiling studies. As a result, many published NMR-based metabolic profiling studies still continue to include putatively identified metabolites and unknown features without providing unequivocal proof

of assignment or they simply label peaks as 'unknown', thereby potentially missing key mechanistic information.

To avoid the problem of multiple entries for the same compound in databases under different names, a community-wide effort is underway to develop better, faster and more standardized metabolite identification strategies, such as implementing standard nomenclature for newly identified metabolites using the International Chemical Identifier (InChI)[17]. Sumner et al proposed a four-level system[18] for assigning a confidence level to newly identified metabolites in metabolic profiling studies: i) positively identified compounds (with a name, a known structure, a CAS number or an InChI identifier), ii) putatively annotated compounds using spectral similarity with databases but without chemical reference standard, iii) putatively identified chemicals within a compound class, and iv) unknown compounds. Wishart et al. proposed a further distinction for those metabolites: the "known unknowns" and the "unknown unknowns"[19]. A "known unknown" corresponds to a metabolite that has not yet been identified in the sample of interest but that has been previously described in a database or in the literature, whereas, a truly novel compound, "an unknown unknown," has never been described or formally identified.

Commercial packages, such as Bruker's AMIX TM software and open-source software[20], such as COLMAR (http://spinportal.magnet.fsu.edu/), can help with identifying these 'known unknowns' and some of these software applications are capable of automatically or semi-automatically annotating a limited number of compounds in a biological sample. However, even with automated annotation, the software still requires manual revision and can be prone to inconsistent interpretation and assignment by different individuals[19]. Most software packages and databases do not support identification of 'unknown unknowns', although a few platforms, such as AMIX, include prediction software to aid the identification of novel compounds.

Open access databases have been created for researchers to deposit information relating to newly identified compounds. Most of the available databases such as the human metabolome database, (HMDB)[21], the BioMagResBank (BMRB)[22], PRIMe server,[23] COLMAR ^{1}H(^{13}C)-TOCCATA, and Bruker-AMIX(http://www.bruker-biospin.com/amix.html) contain chemical shift values, relative intensity and peak shape information for ^{1}H-NMR and often ^{13}C-NMR data to support metabolite identification. However, all databases contain inherent errors such as incorrect structures for the metabolites, incorrect names and incorrect assigments. This problem is compounded further by the impact that experimental conditions such as the pH or ionic content of the sample can have on the chemical shift of a

metabolite. Some of these databases, such as HMDB, provide complementary information, including mass spectrometry assignments, which can be useful for checking potential errors in assignments of NMR peaks.

However, although there are resources available to aid assignment of candidate biomarkers, there is no panacea for accurate metabolite identification and there remains a clear unmet need for improved strategies for metabolite identification and curation for NMR spectral profiling.

NMR spectroscopy for structure elucidation

NMR spectroscopy exploits the molecular property of spin and the fact that small differences in the local electronic environment of a molecule will result in differences in properties such as chemical shifts that relate to specific chemical groups. NMR is the only spectroscopic tool that delivers atom-centered connectivities information, giving it a major role in molecular structural elucidation. It has many benefits over the other techniques currently employed in metabolic profiling, such as minimum sample preparation, high reproducibility and robustness, non-destructive nature and relative quantification without using internal standards[24]. Additionally, although it is generally less sensitive than MS-based spectroscopic platforms, NMR is capable of providing wide selectivity with respect to analytes and definitive structural information for detecting them with no restrictions on polarity, volatility, or chromophore content[25]. However, development of magnets with increased field strength, operating frequencies and cryogenically cooled probes are enhancing the sensitivity offered by NMR. Typically, NMR requires larger volumes of samples for high-throughput profiling studies (0.3–0.5 mL) than some of the MS-based methods, but where high throughput is not a prerequisite of the study design, smaller volumes (30 µl) can be measured using microtube technology[26].

The key advantage of NMR spectroscopy in metabolic profiling lies in its exceptional reproducibility and ability to exploit atom-centred properties, making structural elucidation of chemicals relatively easy. Newer generation NMR instrumentation and technology, such as the IVDr standardized platform for NMR biofluid screening, allows an effective, reproducible, and high-throughput approach to metabolic profiling as applied to clinical diagnostics or food quality control. Standardized protocols that minimize technical or experimental bias have been reported in the literature for urine, blood[27], tissue extracts[13] and NMR-based plant metabolite profiling analysis[28]. If the chemical environment (temperature, pH and ionic strength of the solution) is well-controlled and standardized protocols[13,27] are followed

rigorously, comparable data sets can be generated from different laboratories. The success of large-cohort studies of metabolic profiling of biofluid samples with minimal batch effects has made NMR the first choice over other analytical techniques for large-scale population screening that does not require batch correction or statistical manipulation.

In metabolic profiling studies, multivariate statistical analysis is typically used to interrogate biological spectra relating to a biological metric or class. These metrics can be continuous variables such as blood pressure values or relative quality of a food product, or can be binary classifiers such as presence/absence of a specific disease. Identification of metabolites or signals associated with the biological metric is often the aim of the study and forms the basis of biomarker discovery, although it is likely that at least a portion of the signals identified as being associated with a particular biological property will not be known. The first dataset acquired is 1D proton NMR (^{1}H-NMR) data, which contain information that can be utilised in the identification of signals such as: chemical shift; multiplicity and shape of the signal; the homonuclear (^{1}H-^{1}H) coupling constants; the half band-width of the signal; and stability and integral of the signal (intensity ratio of the signals from the same compound). Although tentative assignments can be made using these spectral properties, they are almost never sufficient to accurately and unequivocally assign the chemical compounds.

To access more chemical information, 2D pulse sequences can be used to disperse the signal into a second dimension to increase resolution and tackle the problem of feature overlap. Homonuclear and heteronuclear experiments-such as ^{1}H−^{1}H *J*-resolved (*J*res)[29-32], ^{1}H-^{1}H COrrelation SpectroscopY (COSY)[14], ^{1}H-^{1}H TOtal Correlation SpectroscopY (TOCSY)[14,33], ^{1}H-^{13}C Heteronuclear Single-Quantum Coherence (HSQC) with and without multiplicity editing[14] and ^{1}H-^{13}C Heteronuclear Multiple Bond-Correlation (HMBC)[14] spectroscopy are used to access information relating to coupling constants, ^{1}H-^{1}H correlation or ^{1}H-^{13}C correlations (Table 1).

1D NMR experiments can also be used for structural elucidation. For example, selective TOCSY (sel-TOCSY)[34] might address the feature overlap issue to facilitate signal assignment of coupled protons along the same unbroken network of couplings. 1D sel-TOCSY is more specific for the resonance(s) in question from the unknown metabolite and takes less time than 2D TOCSY. This experiment may also provide the multiplicity of the resonances within the network and may match with what is observed on 1D 1H spectrum, but this will depend on the spinlock mixing time. Either way, the resulting multiplicity may indicate what type of spins are attached to the fragment from where the resonance was

selectively irradiated. This experiment can also be useful for tackling, to some extent, dynamic range issues. For instance, correlations between metabolites that are present at lower concentrations and whose resonances network is observed on both aromatic and aliphatic regions, which can be obscured on 2D TOCSY by resonances from highly concentrated metabolites in the same sample, such as sugars in food, urine, serum and plasma. By irradiating the resonance on the aromatic region, it is possible to observe the resonances on the aliphatic region and confirm peak assignment. We recommend running this experiment on sample(s) where the targeted resonance(s) are less overlapping and performing additional experiments in reverse direction.

^{13}C Distortionless Enhancement by Polarization Transfer (DEPT) provides information about the type of protonated carbons in a spectral editing fashion. This experiment can be acquired after a given purification method has been applied to a targeted sample. In cases where there are multiple independent spectra acquired for a particular biological class or condition, use of the statistical correlation between signals originating from the same molecule may obviate the need for further spectroscopy experiments.

Where the molecule cannot be identified by NMR experiments alone (e.g. due to the limited amount of sample available or NMR providing insufficient information to characterise consclusively), then hyphenated chromatograpic-mass spectrometry techniques can be used to facilitate the identification of unknown compounds by increasing sensitivity and simplifying the laborious and complex traditional sample purification methods. However, the isolation of compounds remains challenging when they are present in low concentrations or in complex biological matrices. The development of directly coupled HPLC-NMR[35], HPLC-NMR-MS[16] and LC-SPE-NMR[15] approaches is targeted towards achieving unambiguous chemical characterization of endogenous and exogenous metabolites.

Statistical tools to enhance molecular structural elucidation

Statistical correlation methods can be used to investigate whether two peaks belong to the same molecule using spectral data already acquired, provided there are sufficient samples in a given biological class to perform a correlation analysis. The ratio between peak intensities/integrals in ^{1}H-NMR spectra is directly related to the chemical and molecular properties (number of protons of each multiplet) of a molecule. The ratios of a molecule are constant across all spectra, so the integral of a peak is directly related to the concentration of the molecule in a sample. The correlation between spectroscopic variables can be used to identify structure within the data, where high correlations indicate a high likelihood of two

peaks having the same ratio. A low correlation indicates the ratios of the peaks are different across different spectra and they are likely to belong to different molecules[36]. This type of analysis, which is commonly referred to as statistical spectroscopy, allows recovery of structural and pathway information from analysis of sequential or parallel spectroscopic measurements on multiple samples[37]. The concept that correlative structure in spectra could be exploited to extract chemical information was first applied to Raman and infrared spectra[36]. The use of the Pearson or Spearman correlation to identify structure in ^{1}H-NMR spectra was subsequently developed by Cloarec[9] et al. in 2005. Use of previously acquired spectral data instead of conducting further analytical experiments results in a substantial reduction in both cost and time. For example, an HSQC experiment on a single urine or plasma sample typically takes around 7-8 hours, whereas the statistical correlation will take in the order of seconds to minutes depending on the size of the dataset. However, statistical methods are no replacement for actual analytical experiments, but rather serve to guide analytical experiments in a targeted fashion [11].

Statistical TOtal Correlation SpectroscopY (STOCSY) – One of the characteristics of NMR spectroscopy is that multiple peaks in an NMR spectrum can derive from the same molecule, and irrespective of metabolites' concentration, they always appear in the same relative intensities to each other. This can be exploited by calculating the correlations between variables in a spectral dataset, where high correlations indicate that these variables (clusters of spectral data points) likely belong to the same molecule. This approach[9] is commonly used in metabolite identification for NMR metabolic phenotyping. STOCSY gets its name from the TOCSY experiment, which identifies coupled spins from all protons within a given spin system. A TOCSY spectrum is a 2D representation of correlations between coupled protons within a single sample. STOCSY, on the other hand, uses multiple 1D ^{1}H-NMR spectra to calculate statistical correlations between all data points in the spectra, so it works on all spin systems of a molecule and is not affected by the distance between any two protons. A 2D pseudo spectrum is created that resembles a TOCSY, but shows spectrum correlations between all protons of a molecule (not just the coupled protons). Instead of calculating correlations between all data points, the 1D version of STOCSY creates a pseudo-spectrum that shows the correlation between a single data point (driver variable) and all other variables. It is equivalent to extracting the correlations of a single row/column from a 2D STOCSY. 1D STOCSY visualizes the analysis by showing the covariance between the driver variable and all other variables on the y-axis as a function of the chemical shift on the x-axis. This pseudo spectrum resembles a real ^{1}H-NMR spectrum in that it preserves the ratio between peaks and multiplets, and simultaneously indicates the sign of correspondence of the driver with other peaks. The (absolute) correlation of each variable

with the driver is then indicated using a colour scheme with a gradient to show the magnitude of the correlation (positive: similar ratio; close-to-zero: no relationship between ratios; negative: inversely similar ratio). Variables with high correlation and possibly high covariance likely belong to the same metabolite. However, where there is a degree of overlap of signals in a particular spectral region, statistical associations may be reduced.

SubseT Optimization by Reference Matching (STORM) – Since each NMR spectrum consists of hundreds of peaks that can partially overlap, and will do so differentially in different regions of the spectra, lower correlations between variables may result if other chemical species confound the association. In order to reduce the effect of peak overlap, a subset of spectra in which the peak of interest is more clearly visible can be used instead of the entire data set. STORM[10] uses a reference signal to evaluate which spectra contain signals that resemble the peak shape in the original reference signal. The reference signal is chosen based on prior knowledge of the potential identity of a peak. For example, the reference signal can be the peak shape from a pure standard of a known molecule, but more commonly it is observed in the data that do not overlap with other peaks in that spectrum. This happens when the identity of a molecule is unknown, but there is information on the peak shape that is of interest (e.g. a specific multiplet pattern was observed). All spectra that do not contain signals that resemble the peak shape of the reference signal are excluded from the analysis as they would only contribute 'noise' to the correlation (as the concentration of the target metabolite is likely below the detection limit or distorted by overlap with other signals). Visualizing the correlations from the 'clean' subset in the same way as STOCSY gives a clearer description of structural correlations in the data set. The second aspect incorporated into the STORM algorithm is the use of multiple testing techniques to lessen the possibility of reporting false positive associations. STOCSY and other related methods[37] such as iSTOCSY, CLASSY, SRV solely focus on 'high' correlations, which does not account for sample size, instead of focussing on significant p-values, which do depend on the sample size. STORM uses different methods, such as the Bonferroni correction, to control the Family-Wise Error Rate and methods such as the Q statistic or Benjamini Hochberg correction to control the False Discovery Rate[38,39].

RED-STORM – The rationale behind STORM relies on the assumption that there are a number of spectra that have a clear spectral signature of the unknown metabolite. However, it is possible that other peaks from a metabolite appear in regions of the spectrum where many common metabolites have signals at the same chemical shifts. In these cases, it will not be possible to use STORM to identify all peaks and spin systems from a metabolite. Therefore, the concept behind STORM was extended to 2D NMR experiments such as 2D *J*-

resolved spectroscopy. In these experiments, the pulse sequence separates peaks from the same multiplet along an orthogonal axis that reflects the *J*-couplings. STORM was adapted to operate on 2D J-resolved data sets, and the resulting calculation was put into a Bayesian framework to achieve subset selection and assess variable importance. This method was termed Resolution EnhanceD (RED-) STORM[11] due to the extra dimension giving the increased resolution needed to separate overlapping peaks.

Overview of our system

Here, we describe our system for identifying both known and unknown metabolites derived from NMR metabolic phenotyping studies of complex biological samples (such as urine, serum, plasma, faeces, multiple tissue extracts, breast milk and food or plant extracts). This system is the result of previously published strategies that have successfully identified metabolites[11,40-45]. It includes 8 workflows based on advanced statistical and analytical methods that are organized according to the level of difficulty and time requirements (Figure 1). These workflows can also be applied to validate the identification of compounds through automatic annotation and to expand the list of "known unknowns" in the increasing number of large-scale studies that utilize a targeted analytical approach.

In the standard approach, the samples are prepared and analysed, data are acquired (using 1D or 2D NMR) and analysed (typically using STOCSY to begin with), and then these data are matched with those in the relevant databases to identify biomarkers (Figure 1A). Computational modelling of the NMR data is performed against a classifier (e.g. disease versus control) or a continuous biological response metric (e.g. body mass index [BMI] or serum liver enzyme level). Typically, multivariate linear projection methods (such as principal components analysis (PCA) or projection to latent structures discriminant analysis (PLS-DA)) are used to generate a classification or predictive model in which the loadings or coefficients of the components or vectors can be identified and related to signals from molecules that are characteristic for a specified condition. These methods can be combined with orthogonal signal correction for removing/extracting[46] hidden structure in the data[47] Other methods of identifying candidate molecular signatures for a given physiological or pathological variable include machine learning techniques such as random forests[48,49] and genetic algorithms[50,51] or clustering techniques such as K-nearest neighbour[52,53] or hierarchical cluster analysis[54,55].

An NMR peak can be used to discriminate between two or more biological variables if– after performing any of the methods above – it meets both of the following requirements: (1) it is statistically significant after adjustment for multiple testing using either Family-Wide Error Rate (FWER) adjustments (e.g. Bonferroni correction) or False Discovery Rates (e.g.

Benjamini-Hochberg, Benjamini-Yekutieli or Storey-Tibshirani/Q-value FDRs) and (2) it is adjacent to a 'significantly' discriminatory datapoint on either side, with the same sign as both adjacent data points. FWER methods are too strict because they aim to prevent making one (or more) type-I errors, whereas FDR methods aim to control the type-I errors at a constant level (e.g. 5%) for the entire dataset, which is beneficial given the inherent correlation structure of the data. The Q-value (Storey-Tibshirani FDR) approach gives a direct estimate for each individual variable as to what proportion of false positive discoveries (type-I errors) are made for all variables that are 'as significant or more' (lower or equal *P*-values) and it is able to deal with large numbers of true positives and some form of dependence (correlation) between them. Sometimes the chemical shift and multiplicity of the 'biomarker' signal can be matched against existing databases containing structural information and a positive identification can be obtained without further analysis. However, to efficiently increase the amount of information available for matching against the database, we recommend performing STOCSY (Fig. 1b, workflow 1) before searching databases.

Alternatively, it is possible to proceed directly to hypothesis driven methods such as STORM (Fi. 1c, workflow 2) and RED-STORM (Fig.1c, workflow 3) if the computational modelling of the NMR data and the information about the multiplicity of the unknown is provided, as this is essential to enable selection of the best subset of samples on which to perform STORM and RED-STORM . It is also possible to proceed directly to STORM in cases where the unknown signal is small or hidden in the baseline. Unless the 2D J-Res spectra have been previously acquired for each sample (which may not be always feasible due to time or financial constraints), we recommend that the user does not proceed directly to RED-STORM .

Workflow 1 (STOCSY)

Where there are sufficient spectral objects in each of the classes, STOCSY can be used to identify other signals in the NMR spectra relating to the same molecule. The premise of this approach is that datapoints from the same molecule will generally have a stronger correlation with each other than with those from unrelated molecules, noise or from molecules in the same biological pathway, which may also be correlated (but generally to a lesser extent). This principle does not always hold true in cases where the signal is present in a crowded region of the NMR spectrum where there are overlapping signals from other molecules. The correlation threshold is set to determine whether two correlated signals arise from the same molecule (structural correlation) or from molecules in shared biochemical pathways or subject to shared biological phenomena[56,57] (mechanistic correlation). The code

for STOCSY can be found in the STORM algorithm (https://bitbucket.org/jmp111/storm/src). The algorithm is set up to conduct either STOCSY (workflow 1) or STORM (workflow 2). To generate a simple pseudo-spectrum showing correlation values (i.e. the STOCSY method), the STORM algorithm should be run using the command JMP_STORM (ppm,X,driver_ppm) where ppm_driver is a single datapoint entry for the vector with chemical shift values. Other software packages such as SIMCA-P and R (e.g. MWASTools package from Bioconductor[58]) also have capability for carrying out the STOCSY procedure.

To achieve the best results, the datapoint acting as the 'driver' for the correlation coefficient should be the point in any cluster of datapoints in the coefficient plot (either regression, correlation or other metric) showing a high association with the class (e.g. a continuous measure such as BMI or binary case-control status), provided that it is bound on either side by at least one datapoint that is also significantly associated with class[40]. In most cases, the most significant datapoint for any given peak will be at the apex. However, if there is peak overlap, then this will not necessarily be the case. Sometimes the most statistically significant peak is partially obscured by a larger signal that shows no co-variation with the model classifier.

The full set of correlated resonances identified using the STOCSY algorithm, and multiplicities of those signals and intensities, can then be checked against existing spectral databases followed by standard spiking. In cases where no correlations are found, a more stringent statistical workflow such as STORM (workflow 2) can be used**.** It is important to highlight that a variation on 1D STOCSY is the use of a second nucleus (e.g. ^{13}C or ^{31}P) to establish statistical connections between protons and the second nucleus[59]. In a study exploring the metabolism of fluorinated drugs, the ^{19}F nucleus was used to simplify the complexity of overlapping drug-related and endogenous peaks in the proton spectra of urine samples of individuals after the administration of flucloxacillin[60].

Should STOCSY followed by database matching and standard spiking provide unconclusive results, try the alternative workflows outlined in Fig.1c (workflows 2-8). Please bear in mind that they are outlined according to levels of difficulty, but it is possible to change their order according to the type of unknown, amount of sample and resources available. More detail about each of these possible variants can be found below. The first two variants (workflows 2 and 3) do not involve acquiring new data. Instead, the existing data is subjected to a different type of statistical spectroscopy – STORM (workflow 2) and RED-STORM (workflow 3).

Matching data to databases

The chemical shift, multiplicity and relative intensity information is then matched against NMR databases of chemical standards. Open-source databases, including the human metabolome database (HMDB; http://www.hmdb.ca) and the Biological Magnetic Resonance Data Bank (BMRB; http://www.bmrb.wisc.edu), contain NMR spectra data. HMDB also contains mass spectra data from chemicals typically found in human biofluids. These databases can be searched by chemical name, formula or by chemical shift. In addition to open source databases, there are commercially available databases such as Chemomx (https://www.chenomx.com/software/libraries/) and Bruker BBIOREFCODE (https://www.bruker.com/products/mr/nmr/nmr-software/nmr-software/bbiorefcode/overview.html).

Where a 'good' match is found between the candidate biomarker and a database assignment, then the compound will be putatively identified based on spectral similarity with databases or putatively identified within a compound class. However, even if biofluid samples are stabilized by a buffer, the pH can differ due to precipitation of chemicals over time causing chemical shift instability. Additionally, differences in divalent metal ion concentrations can incur inter-sample differences in chemical shift[61], which can make matching by chemical shift alone unreliable. Also, it must be noted that even the largest databases do not comprehensively cover the breadth of molecules present in biological samples.

Spiking chemical standards

In order to positively identify the compound, the incremental spiking of authentic chemical standard should be carried out. In the case of identifying an unknown unknown, the name, structure, CAS number and or International Chemical Identifier should be deposited in a database. If the standards are not commercially available, or it is not feasible to synthesise the compound, it should remain as a putatively annotated compound or putatively identified within a compound class.

Once a candidate chemical structure is identified for the unknown metabolite, the chemical standard is acquired and dissolved in phosphate buffer (see reagents set up and sample preparation sections). There are two strategies for spiking of synthetic or commercial standards into biological samples. The standard is made up at a specific concentration representative of 1.5 to 2 times the concentration of the candidate peak in the biological sample. An initial volume of 2.5 µl of the standard is added to the biological sample and a 1D NMR spectrum is acquired. This process is repeated a further 3-5 times in increments of 2.5

µl giving a series of NMR spectra with incremental concentrations of the candidate compound (Fig. 2h). If the candidate is assigned correctly, all peaks belonging to a given molecule should increase 'cleanly' in intensity in the correct relative ratio. It is possible to spike higher volumes but this will result in sample dilution which may affect spectral quality. We recommend adding a set volume of increasing concentrations of sample e.g. 2, 4, 6 millimoles dissolved in a volume of 2.5 µl (Fig. 5 g and h).

Where the chemical standard is not commercially available, then chemical synthesis is required. If this is still not possible, the compound will remain as putatively identified. If the compound is positively identified, we recommend that the standard spectrum and spectral parameters are incorporated in a spectral database.

Workflow 2 (STORM)

STORM (a statistical modification of STOCSY) should be performed in cases where STOCSY fails to detect other correlated signals, or where spectral overlap reduces the ability to detect systematic, simultaneous variation between two or more signals. By iteratively modelling correlation from the apex of a driver signal and subsequently reducing the dataset to include only those spectra with signals relating to the 'unknown' metabolite, a clearer picture of the spectral structure of that peak can be obtained[10]. This iterative approach aims to find a subset of spectra that contain the spectral signature of the 'unknown' molecule. A critical aspect of this workflow is to select a good reference for the 'unknown' molecule. For instance, if the pattern of positive (negative) regression coefficients has two local maxima (minima), but the average of the spectra only shows one peak, the reference should include two peaks at the exact locations of the local maxima (minima) instead of a spectrum that resembles the average or majority of spectra.

A key limitation of this workflow is that local misalignments of the signal of interest (due to pH or variable presence of divalent metal ions) may result in a reduced spectral subset, as not all spectra containing the metabolite signals may be selected for particular modelling steps. However, locally clustering the data based on the correlation structure and selecting one spectrum to serve as a reference for each cluster can identify all subsets of spectra containing the unknown molecule of interest.

The algorithm also allows for inclusion of multiple regions of interest. However, in order to calculate local correlations between spectra, and hence identify a subset of samples containing the target molecule, it requires a reference that contains at least three adjacent variables. Where multiple correlated signals are detected, the chemical shift, and optimally recovered peak shapes of the signals from the 'unknown' molecule can then be matched

against the spectral databases, followed by spiking of authentic chemical to positively identify the compound.

Workflow 3 (RED-STORM)

It is increasingly common to acquire a 2D J-Res spectrum of each sample directly after the standard 1D spectrum as this adds only a short increase in run time per sample (typically 7 mins). Where J-Res spectra are available, a 2D extension of the STORM algorithm (workflow 2) can be exploited. *J*-Resolved experiments provide information on the multiplicity of signals by adding a second dimension that represents the proton-proton coupling, with the added advantage that the 1D projection is a decoupled proton spectrum that simplifies interpretation by reducing the peak overlap[62]. This 2-D spectrum requires a relatively short acquisition time but can be limited in characterising molecules with short T_2 and by artefacts introduced by second order coupling. However, these artefacts may be used for identification purposes as after symmetrization along f1 they often appear in the middle of signals from the corresponding coupled protons within a second-order system. These artefacts can be observed as “stronger couplings“ than their parents in terms of Hertz in the second dimension. For instance, the multiplet observed at ~ 3.75 ppm from ascorbic acid[10] (Fig. 2e) is derived from two doublet of doublets with an artefact in the middle at ~ 4.03 ppm. Variations on the basic *J*-Resolved pulse sequence have been proposed for use in metabolic profiling[30]. We recommend following the *J*-Res workflow for subsequent multivariate modelling developed by Dona et al[27]. Where multiple correlated signals are detected from the *J*-Res spectra, the correlated peaks can be matched against a spectral database. If neither STORM nor RED-STORM is successful, analyse the sample using a different NMR method such as COSY or HSQC (workflow 4), ideally using a fresh sample.

Workflow 4 (conventional 2-D NMR spectroscopic analysis)

Conventional 2D NMR spectroscopic analysis (workflow 4) can be used to extract more information about molecular structure when STOCSY, STORM or RED-STORM do not identify a plausible metabolite. A representative sample with high concentrations of the unidentified metabolite is selected and measured using one or more of the pulse sequences below:

i) Homonuclear correlation spectroscopy (e.g. Correlation SpectroscopY (COSY)): This pulse sequence allows the measurement of spin-spin coupling up to 4 bonds away, but mainly detects neighbouring protons[63]. TOtal Correlation SpectroscopY (TOCSY) measures spin-spin coupling of up to 6 or 7 bonds distance depending

on the length of the spin-lock time in the pulse sequence and displays proton-proton couplings on the off-diagonal, allowing the connections between adjacent protons to be reconstructed[64]. The magnetization transfer is disrupted by the presence of heteroatoms such as oxygen or nitrogen, quaternary carbons and carbonyl group.

ii) Heteronuclear spectroscopy (e.g. HSQC): In this experiment, direct coupling between protons and a second nucleus with spin I=1/2, usually ^{13}C, are observed, using the higher sensitivity of the proton to observe the lower sensitivity nucleus through direct coupling[65,66]. The HMBC pulse sequence detects ^{1}H-^{13}C connections across multiple (typically up to four) bonds[67].

Acquisition parameters for each of these experiments are set out in Dona et al[27] (standard 1D experiment, Carr-Purcell Meiboom-Gill (CPMG), *J*-resolved) and in Lindon et al[68] for the remaining 2D pulse sequences. A standard 1D spectrum of each sample should be reacquired prior to applying a 2D pulse sequence, as well as an additional 1D spectrum at the end of the series of 2D experiments to assess sample integrity, as typical 2D spectra take in the order of a few hours to acquire. In order to obtain good quality 2D spectra within an acceptable timeframe, the number of scans and experiments for the second dimension (also known as increments) can be selected on the basis of the signal intensity of the unkown resonances that can be compared to the TSP singlet. For instance, if the signal intensity is very small (i.e. the metabolite is diluted), it is advisable to run higher number of scans than increments.

Other 2D NMR experiments that provide additional structural information include diffusion-ordered experiments (DOSY), which capture information about molecular tumbling times and hence molecular mobility and size[69]. Information about molecular size and mobility can also be obtained through use of diffusion edited pulse sequences[70]. 2D DOSY (Difussion Ordered SpectrocopY) provides a 2D spectrum with chemical shift on the horizontal domain and diffusion coefficient (D in m^2/s, normally expressed as log(D)) on the vertical dimension. Thus, DOSY is able to separate resonances of different metabolites in a second dimension according to their different D values, which depend on molecular properties such as size, shape, mass, charge and mobility. Accordingly, signals from the same metabolite will be observed on the same horizontal axis with its corresponding D value. The information obtained by this type of experiment is similar to that observed on chromatograms but also includes the NMR signals, which makes this experiment useful for confident peak annotations. It is worth noting that this experiment may not work for overlapping signals from metabolites with similar D values. For these cases, chromatography methods should be applied instead. However, in some other cases, DOSY is able to separate and analyse

molecules with the same size and structure and similar D values, such as α- and β- anomers of carbohydrates. As diffusion is also affected by solvent viscosity, temperature, pressure and convection inside the tube, these conditions should be consistent throughout the analysis of all samples included in a study.

Once a set of correlated signals are identified, their combined structure is checked against existing databases or deduced *de novo* and a solution of the chemical standard made up in the same buffer as the sample is spiked into the sample. Sometimes, workflow 4 does not derive conclusive information because the unknown molecule is overlapped by other molecules or present at lower concentations than the NMR limit of detection. If this is the case, please proceed to workflows 5 or alternatively to 6 or 7 according to the amount of sample available. Workflow 5 is less time consuming and more simple than workflows 6 or 7. However, if not enough sample is available to conduct 5 and 6 or 5 and 7, it is recommended to try workflow 6 or 7 first to be able to conduct workflow 4 afterwards if neccesary.

Workflow 5 (concentrate the sample before re-analysis)

In many cases, the discriminatory signals for a particular biological class may be present in low concentrations, or the signals may be overlapped by those from other, higher concentration molecules, making identification difficult. When statistical correlation spectroscopy and standard 2D NMR techniques fail to unequivocally identify a metabolite, the sample should be lyophilised and reconstituted in a lower volume to increase the concentration of the metabolite, typically by 2-10 fold. For lyophilisation, a volume of 5-10ml urine or other biofluid can be dried down and reconstituted in 1ml of water [71]. Then, the appropriate amount of the reconstituted sample should be mixed with the appropriate buffer (as described in PROCEDURE, sample preparation). If the lyophilised sample is reconstituted into a volume that is too small, it may not be possible to completely dissolve the dried sample or it may not be easy to tune the spectrometer due to the high ionic strength of the reconstituted sample.

Moreover, the difference in concentration between the original and reconstituted sample, and the potential difference in sample pH after undergoing elution with acid wash and other solvents during a solid phase extraction process (particularly after undergoing solid phase extraction chromatography), can result in changes in the chemical shift and peak shape of the candidate unknown signal.

The preconcetration of the unknown compound remains challenging when it is present in low concentrations and or in complex matrices. Often this sample concentration step is

performed by simple chromatographic separation workflows such as solid phase extraction chromatography (SPE) (Fig. 5d) or liquid-liquid extraction, which will remove other compounds from the sample matrix and enrich the candidate unknown. Appropriate methods and protocols are outlined in Lenz, 2011[71]. Once the 'unknown' molecule has been concentrated, perform conventional 2D NMR spectroscopic analysis **(worflow 4)**.

Workflow 6 (LC-NMR-MS analysis)

For LC-NMR-MS analysis, a biological sample or previously pre-concentrated sample is injected into the LC column (for example 5ml urine or other biofluid can be dried down and reconstituted in 500ul of either buffer or deuterium oxide). Following the method described by Shockor et al[16], chromatographic separation of the sample will result in fractions collected every 29 s. Each of these fractions should be dried under a stream of nitrogen and reconstituted in 540 ul of water and 60 µl of the phosphate buffer for urine samples (see procedure 2F for LC-fraction collection and 4B for data acquisition). This sample is then analysed by ^{1}H-NMR to identify the fraction with the highest concentration of the unknown candidate. Once the unknown compound is isolated (Fig. 4a), further 2D NMR analysis can be performed on a "pure" sample (Fig. 4d) to generate further structural information that can be complemented by mass Spectrometry analysis as detailed in Fig. 4b,c-e.

The challenge of this workflow is to be able to isolate the unknown signal in one or two fractions. Frequently, several NMR-visible peaks that belong to other compounds will be present in the same fraction. However, this can be solved by changing the gradient in the liquid chromatrography analysis and even modifying the fraction collection times in order to isolate the unknown compound as cleanly as possible. If there is enough sample, it is also possible to run the fraction again using different chromatographic columns and different gradients in order to improve the isolation of the uknown. Even if the unknown compound shares the fraction with other compounds, enhanced 2D-NMR analysis can provide better resolution than the original 2D-NMR analysis, as most compounds (and therefore the peak overlap) will be removed.

Workflow 7 (LC-SPE-NMR-MS analysis)

SPE cartridges can be employed to increase the sensitivity and quality of the NMR measurement over the conventional LC-NMR method. Unknown compounds are eluted from the SPE cartridge to the NMR flow probe using deuterated acetonitrile for initial NMR investigation. Then, the fraction can be recovered by flushing the sample out of the NMR probe head with nitrogen gas and mass spectrometry analysis can be performed on the recovered fraction. A detailed description of the method is set out in Godejohann et al[15].

The main challenge of this workflow, as for the other workflows, is completely isolating the unknown signal. However, even if the unknown compound shares the fraction with other compounds, most compounds and signal overlap will be removed from the original sample. Therefore, enhaced 2D-NMR analysis with better resolution can be obtained. If there is a sufficient sample, it is possible to run the fraction containing the unknown compound using different SPE cartridges in order to improve the isolation.

Workflow 8 (combining multiple analytical platforms)

If the unknown compound has been isolated or almost isolated by workflows 5,6 or 7, further structural identification can be undertaken using one or more analytical platforms, such as CE-MS [7273] and GC-MS[74]. Previously established metabolic profiling methods can be used to take advantage of the strengths in separation and detection of other analytical platforms. In addition, the use of sample derevitization or complex formation can be useful if the molecular structure has partially been elucidated. Once a candidate structure is identified, it is checked against existing databases or deduced *de novo,* and spiking of the putative chemical standard is conducted for confirmation of identify.

Advantage and Limitations of our system

Many of the signals derived from NMR metabolic phenotyping studies belong to molecules that are not present in databases. Identification of these "unknowns" is notably difficult, expensive and time consuming. Despite the fact that NMR spectroscopy has a long track record in the structure elucidation of unknown biological molecules, particularly for natural product research, it requires the isolation of the unknown molecule, which is a time consuming process and some metabolites can be modified or lose their activity during extraction. The combined used of hyphenated chromatographic-mass spectrometry techniques, such as HPLC-NMR[3536,75], HPLC-NMR-MS and LC-SPE-NMR[15] systems, facilitates the identification of unknown compounds. These methods can increase sensitivity and simplify the laborious and complex traditional sample purification methods but they are limited in cases when the unknown compound is present in low concentration.

The integration of data generated from the same cohort of samples by NMR and a second analytical platform has been used for identifying and validating candidate biomarkers through statistical tools such as Statistical HeterospectroscopY (SHY)[76] and bidirectional Pearson correlations between NMR and Capillary Electrophoresis[42] and NMR and Gas Chromatography coupled to Mass Spectrometry[77]. In addition to achieving greater resolution

at the structural level, they provide a higher level of biological information on metabolic pathway activity by examining the different connectivities in both the correlation and anticorrelation matrixes. However, further application of these hyphenated methods to identify 'unknown unknowns' is required and therefore they have not been included in our proposed system.

Finally, molecular networking, whereby metabolites are mapped to biochemical pathways and statistically connected where relevant, has been widely employed in both the MS and the NMR communities to establish relevant biochemical pathways in disease-based studies and to add confidence to metabolite annotations based on over-representation of particular pathways ([78,79]). However, full discussion of network analysis is beyond the scope of the current paper and merits separate consideration, as there are numerous methods with multiple purposes and outcomes.

EXPERIMENTAL DESIGN

To ensure the efficiency of the protocol, it is necessary to give prior consideration to certain aspects of sample collection (**step 1 of the Procedure**), storage, preparation (**step 2 of the Procedure**) and processing (**step3 of the Procedure**) that vary depending on the sample type. For **step 1 of the Procedure**, the use of additives or stabilizers (e.g. EDTA) for serum and plasma should be avoided as these additives will generate signals, which may interfere with endogenous signals in both the ^{1}H and 2D NMR spectrum. Preservatives such as boric acid are often added during urine sample collection to avoid bacterial contamination. Be aware that boric acid forms complexes with endogenous metabolites, such as citrate and α-hydroxyisobutyrate, in the urine sample. In addition, samples with high boric acid amounts may not be compatible with the application of other metabolic profiling methodologies such as GC-MS, which is suggested as part of workflow 8 (Fig.1). Therefore, we suggest that the collected samples are sub-aliqoted, using appropriate containers and appropriate storage conditions (**step 1 of the procedure**) to avoid sample degradation and or contamination. If possible, we recommend the use of an independent aliquoted sample, with no previous freeze-thaw cycles, to carry out the workflows described in Fig.1, as freeze-thaw cycles may affect the quality of the sample. Furthermore, preparation of urine and faecal samples **(step 2A and C of the Procedure)** and processing of tissue, breastmilk and food extracts **(step 3 of the Procedure)** for NMR analysis requires the addition of a phosphate buffer that contains a bacteriostatic agent (sodium azide) to avoid bacterial contamination during data acquisition **(step 4A of the Procedure).**

Finally, the amount of sample available is an important limiting factor to be considered before applying the workflows 3 to 8 (Fig. 1) based on the amounts required for sample

preparation and processing (**step 2 and 3 of the Procedure).** Samples such as urine, faeces, breast milk and food samples are non invasive and therefore it is easier to collect sufficient volumes to perform the workflows suggested in Fig 1., whereas blood samples are invasive and may be difficult to obtain in infants. Tissue samples will be limited by the nature of the procedure and/or the size of the biopsy needle. As it is crucial to have enough sample available to be able to identify compounds that are in low concentration, it is necessary to pre-concentrate the sample prior to carrying out workflows 6 and 7 using strategies described as part of the worflow 5 (Fig.1) such as SPE extraction (**step 2D of the Procedure**) and/or isolation of the compounds using, for example, a LC-fraction collector-NMR-MS (**step 3F of the Procedure**) as described in workflow 6.

MATERIALS

REAGENTS

- Na_2HPO_4, Disodium hydrogen phosphate, 99% anhydrous, (Sigma-Aldrich W239901)
- NaH_2PO_4, 99%, anhydrous (Sigma-Aldrich)
- Perchloric acid (PCA), ACS reagent, 70% wt/wt (Sigma-Aldrich)
- K_2CO_3, ACS reagent, Z99.0% (Sigma-Aldrich)
- D_2O, deuterium oxide, 99 atom % D (Sigma-Aldrich 435767)
- NaN_3, Sodium azide, 99.5% (Sigma-Aldrich S2002) [Caution: sodium azide is highly toxic and highly reactive under certain conditions]
- TSP, 3-trimethyl-silyl-[2,2,3,3-$^2H_4$]propionic acid, sodium salt, 98 atom % D (Sigma-Aldrich 269913)
- Water, HiPerSolv for HPLC, BDH (VWR International Ltd.)
- Chloroform, AnalaR, Z99%, BDH (VWR International Ltd.)
- Acetonitrile, NMR-Chromasolv, Z99.6%, Riedel-de-Ha¸n (Sigma-Aldrich) [Caution:Acetonitrile is highly flammable]
- $CDCl_3$, '100', Z99.96 atom %D, contains 0.03 vol/vol TMS (Sigma-Aldrich)
- Methanol-d4, 99.8% (CD3OD; Goss Scientific)
- Methanol,Sigma-Aldrich 99.8% [Caution:Methanol is highly flammable]
- Formic acid, Sigma-Aldrich [Caution:Formic acid is corrosive and volatile]
- Leucine enkephalin acetate salt hydrate ,Sigma-Aldrich or alternative lock mass compound according to manufacter
- Sodium formate or alternative calibration solution according to manufacturer

EQUIPMENT

- Typically a 600 MHz Avance III NMR spectrometer (Bruker Biospin Ltd.) or similar.
- NMR detector-BBI 600 MHz 5-mm z-gradient probe (Bruker Biospin Ltd) and automated tuning and matching (ATMA) unit (Bruker Biospin Ltd) or similar;
- SampleJet with sample cooling and preheating station
- Topspin 3.5 software with Icon NMR (Bruker Biospin Ltd.)
- Microplate 96 square well 2ml (Fisher Scientific UK Ltd)
- 96 NMR 5 mm tube SampleJet boxes
- Eppendorf tubes, 1.5 ml (VWR international)
- (Bruker Biospin Ltd. Part No. Z105684); POM balls to seal sample tube caps (Bruker Biospin Ltd. Part No. Z72497)
- Typically a Waters Acquity Ultra Performance LC system omprising a binary solvent manager and photodiode array detector with a Waters CTC autosampler with 100 µl sampling needle or similar.
- Xevo G2 Q-TOF mass spectrometer (Waters Ltd), or similar
- Waters Fraction Collector III or similar
- HPLC-column
- Solid Phase Extraction SPE system (Sigma-Aldrich), or similar

BIOLOGICAL MATERIALS

- Urine samples

  [Caution: patient's written, informed, consent should be obtained as well as ethical approval according to relevant Institutional and National regulations].
- Serum samples
  [Caution: patient's written, informed, consent should be obtained as well as ethical approval according to relevant Institutional and National regulations].
- Plasma samples
  [Caution: patient's written, informed, consent should be obtained as well as ethical approval according to relevant Institutional and National regulations].
- Breast milk samples
  [Caution: Mother's written, informed, consent should be obtained as well as ethical approval according to relevant Institutional and National regulations].
- Tissue samples

  [Caution: patient's written, informed, consent should be obtained as well as ethical approval according to relevant Institutional and National regulations].
- food mixtures samples
- faecal samples

  [Caution: patient's written, informed, consent should be obtained as well as ethical approval according to relevant Institutional and National regulations].

REAGENT SETUP

Phosphate buffer for urine samples: As previously described[27], prepare the 7.4 pH buffer by dissolving 20.4 g of KH_2PO_4 in 80 ml of D_2O. Dissolve 100 mg of TSP and 13 mg of NaN_3 (bacteriostatic) in 6 to 10 ml of D_2O. Mix both solutions very well using sonication but be aware that the solution might appear cloudy; this cloudiness should disappear when the pH is adjusted. Adjust the pH to 7.4 by adding KOH pellets. Transfer the solution to a 100 ml volumetric flask and adjust the volume with D_2O. Shake thoroughly to mix completely, and recheck the pH. This volume of buffer is enough to prepare approximately 1500 samples.

Note: This solution should preferably prepared fresh but it can be stored in the fridge at 4°C for a year.

Phosphate buffer for serum or plasma samples: As previously described[27], prepare the 7.4 pH buffer by dissolving 5.32 g of NaH_2PO_4 in 380 mL of water. Add 0.4 g of TSP and shake until the powder is dissolved, add 5 ml of 4% NaN_3 aqueous solution and shake. Add 100 ml of D_2O, adjust the pH to 7.4 by adding 1M HCl/NaOH solutions and fill up the volumetric flask up to 500ml with water. This volume of buffer is enough to prepare approximately 1400 samples. Note: This solution should preferably prepared fresh but it can be stored in the fridge 4-6 ^oC for up to 12 months.

Phosphate buffer for food, breast milk, tissue and faecal samples: As previously described for tissue[13], food samples[11,41] and faecal samples,[80] prepare the pH 7.4 phosphate buffer by weighing 2.62 g NaH_2PO_4, 14.43 g Na_2HPO_4, 1 mM TSP and 3 mM NaN_3 into a 500 ml volumetric flask. Add 100 ml of D_2O and fill up the volumetric flask up to 500ml with water. Shake the flask thoroughly, and leave in a sonicator at 40°C, alternated by shaking the flask, until the salts are dissolved. This volume buffer is enough to prepare approximately 926 samples. Note: This solution should preferably prepared fresh but it can be stored in the fridge 4-6 ^oC for up to 12 months.

LC–MS mobile phases: Prepare enough volume of two mobile phases prior to analysis. The composition of the mobile phase depends on the chromatography method to be used.

Mobile phase A : 100% high-grade water with 0.1% formic acid or similar.

Mobile phase B: 100% methanol with 0.1% formic acid or 100% acetonitrile with 0.1% formic acid. Note: This solution should preferably prepared fresh but it can be stored in the fridge for up to 12 months. [Caution: All solutions should be prepared in a fume hood.]

Leucine enkephalin lock mass solution: Prepare a final concentration of 200 pg μl^{-1} solution of leucine enkephalin in water:acetonitrile (50:50) and dilute appropriately for

positive and negative ionization modes according to the manufacturer's specifications. Store solution at 4 °C until use.

Notes: This solution should preferably be prepared fresh. Alternative compounds can be used for the lock mass solution. Please follow the manufacturer's guidelines for desired concentration.

Sodium formate calibration solution: Prepare a final concentration of 0.1 mg ml^{-1} stock solution of sodium formate in water. Add 1 ml of stock solution to 9 ml isopropanol to give 0.01 mg ml^{-1} solution in 90% isopropanol and 10% water. Solution can be stored at 4 °C until use. Alternative calibration solution can be used according to manufacturer's guidelines.

Note: This solution should preferably be prepared fresh. Alternative compounds can be used for the calibration solution. Please follow the manufacturer's guidelines for deired concentration.

EQUIPMENT SETUP

General NMR setup: Choose the pulse sequence according to the experiment required. Calibrate the instrument at a constant temperature according to the recommendation of the manufacturers for a given biofluid, to avoid modification of small molecules or binding to macromolecules or aggregation of large protein, lipid and lipoproteins. For example, it is usual to acquire serum/plasma spectra at 310K and urine at 300K.

Prior to sample acquisition, calibrate the spectrometer using deuterated methanol[81] (MeOD). Tune and match the probe, and lock to deuterated methanol automatically.

Perform regular shimming followed by automated shimming using a tuning routine. Use a standard 90° proton parameter set to run an experiment with two scans using a pulse length of 1 µs. For all biological samples, suppress the water signal in order to avoid dynamic range problems. Use a typical standard 2 mM sucrose sample (containing 0.5 mM TSP, 2 mM NaN_3 in 90% H_2O:10% D_2O) to check the performance of the water suppression functionality, which is then subsequently evaluated. For Bruker AVANCE III NMR spectrometers, the details of the setup can be found in Dona et al[27] (standard one-dimensional pulse sequence (noesygppr1d), CPMG (cpmgpr1d), *J*-Res (jresgpprqf). For the COSY sequence, a typical acquisition parameter set is given in Holmes et al.[82] For TOCSY and HSQC, see Duarte et al;[83] for HMQC, see Bollard et al[68]; and for HMBC, see Maaheimo et al[84].

Equipment used for solid phase extraction: Condition the cartridges filled with different types of stationary phase (such as *NH_2, SCX, SAX, WCX,* or *WAX* for ion exchange –; and C18, Diol, *CN,or* HLB for reverse phase cartridges) with a solvent prior to loading the

biological matrix in solution onto the column. Chose the stationary phase according to desirable chemical properties such as polarity (e.g. designed to specifically retain anionic compounds or hydrophobic compounds). Elute the biological matrix using a series of different solvents / solvent mixtures, taking into account the nature of the analyte, analyte concentration and sample volume[71].

Fraction collector set up: Connect the fraction collector to the LC system prior to analysis. Place the glass tubes in the tube rack. Follow manufacturer guidelines to set up the starting time, waiting time and frequency for fraction collection according to the LC flow rate and chromatographic method.

General LC-MS set up: Install the appropriate LC column for analysis depending the type of sample and the nature of analytes to be separated. Prepare the mobile phases according to the column and the chromatographic method as described in REAGENT SET UP. Then prime the system and tubing, and condition the column. At the beginning of the analysis, carry out calibration of the mass spectrometer according to manufacturer guidelines. Perform additional instrument system checks if required, according to manufacturer guidelines.

CA-PLS (and PLS, OSC-PLS): The code for executing the PLS, covariate-adjusted (O)PLS and simple orthogonal PLS/PLS-DA is provided in https://bitbucket.org/jmp111/capls/src/. This can be executed in a Matlab environment.

STORM (and STOCSY): The code for executing both the STOCSY and STORM algorithms is in https://bitbucket.org/jmp111/storm/src. These can be executed in a Matlab environment.

RED-STORM - The code for executing the RED-STORM algorithm is provided in https://bitbucket.org/jmp111/redstorm/src/. This can be executed in a Matlab environment.

PROCEDURE

Sample Collection and storage

1| Collect urine, serum, plasma, breast milk, tissue, food or faecal samples using the guidelines described in options A, B,C,D,E,F and G respectively.

(A) Urine.

(i) Collect urine (a minimum of 600µl for 5mm NMR tubes), either spot or 24 h collection, into suitable containers.

(ii) Subaliquot into labelled tubes (for example 1ml eppendorv)

(iii) Store at the lowest available temperature (minimum − 20 °C) until sample preparation (step 2).

(B) Serum

(i) Collect spot blood samples (5ml should provide 5 aliquots of 500µl of serum) into standard Gel vacutainer tubes. Preferably, blood serum should be collected into tubes with no additives (red cap BD vacutainer tubes or similar).

(ii) Place the Gel tube horizontally on ice (not in ice). CRITICAL STEP: this step is crucial to prevent red cell lysis and to reduce protease activity.

(iii) Spin the vacutainer® tubes at 3,100xg for 15 min or up to clotting time at 4°C.

(iv) Subaliquot into labelled tubes

(v) Store at − 40 °C until sample preparation (step 2).

(C) Plasma

(i) Collect spot blood samples (5ml should provide 5 aliquots of 500µl of plasma) into lithium heparin vacutainer tubes at 3,100xg for 15 minutes. Avoid EDTA, citrate and other added stabilizers, which generate additional high-intensity NMR signals that may overlap with other molecules in the spectrum.

(ii) Subaliquot into labelled tubes

(iii) Store at − 40 °C until sample preparation (step 2).

(D) Breast milk

(i) Collect milk sample postfeed (at least 1ml is recommended) by manual expression into collection bottles. CRITICAL STEP: It is better to use a sample of hindmilk due to is composition (richer, thicker, higher in fats and calories), which is why we recommend collecting milk after a feed.

(ii) Subaliquot into labelled tubes, for example 5ml eppendorves.

(iii) Immediately store at − 80 °C until sample preparation (step 2)

(iv) Just before use (step 2), defrost samples at room temperature and vortex vigorously to resuspend the milk fat globules.

(E) Tissue

(i) Freeze the tissue sample rapidly after collection at − 80 °C, preferably using liquid nitrogen to immediately stop any enzymatic or chemical reactions.

(ii) Store in a freezer at − 80 °C until sample preparation (step 2).

(F) Food mixtures

(i) Use gloves to collect the raw foods or meals previously prepared.Immediately after collection, blend the samples using a high speed blender (preferably use a professional blender such as Vitamix) in order to obtain homogenous purees.

(ii) Add water if necessary to facilitate the homogenization . The amount of water depends on the type of food or meal sample. For example, add 250 ml of water (equivalent to the volume of the water typically consumed as part of a meal) to a meal sample or just 30-50 ml for small meals such as porridges, or fruits. The higher the density of the food the higher the amount of water to be added but it is recommended to add the minimum amount as possible. In the case of analysing a set of samples it is adviced to add the same amount of water to all of your samples to make them comparable.

(iii) Aliquot at least 1g of sample into 15 ml labelled falcon tubes

(iv) Store immediately at -80 ^{o}C until sample preparation (step 2).

(G) Fecaes

(i) Collect one complete stool specimen in a faeces collector (such as FECOTAINER, AT Medical BV, The Netherlands).

(ii) Put the stool specimen on ice or store in the fridge within 6 hours of being produced

(iii) Homogenise stool with a sterile spatula and aliquot 300mg of stool into each tube and freeze the samples at -80 ^{o}C until the faecal water extraction is conducted (step 2). However, it is recommended to conduct if possible at collection time, the faecal water extraction from a representative amount (∼15 g) of homogenized stool sample, aliquoted and stored at ≤-20 ^{o}C to avoid further freeze−thaw cycles.

Sample preparation for analysis

2| Prepare urine, blood, and faecal samples for NMR using guidelines described in options A, B and C respectively. Prepare tissue, breast milk samples or food samples for NMR using the guidelines described in option D. Prepare urine samples for solid phase extraction chromatography (SPE) (workflow 5) and LC-NMR-MS analysis (workflow 6) followed by NMR using guidelines described in options E and F respectively. Quantities apply to Bruker 5 mm NMR tubes. Adjust the quantities accordingly depending on different vendor requirements and other variations (e.g. tube size)..

(A) Urine sample preparation for NMR. TIMING ~ 1.5-2.5 h per 96-well plate

(i) Remove sample from frozen storage.

(ii) Thaw the samples on the bench at room temperature(25°C) for 1 hour.

(iii) Centrifuge 600 µl of urine sample at 12000g for 5 min at 4 °C.

(iv) Mix 540 µl of urine with 60µl of phosphate buffer (pH 7.4) in an Eppendorf.

(v) Transfer 600 µl into 5 mm diameter NMR tube prior to analysis and proceed to Step 4.

? **TROUBLESHOOTING**

(B) Plasma and serum samples preparation for NMR. TIMING ~ 1.5-2.5 h per 96-well plate

(i) Remove sample from frozen storage.

(ii) Thaw the samples on the bench at room temperature for 1 hour.

(iii) Centrifuge 400 µl of plasma or serum at 12000g for 5 min at 4 °C.

(iv) Mix 400 µl of plasma or serum with 400 µl of phosphate buffer (pH 7.4) in an Eppendorf.

(v) Transfer 600 µl into 5 mm diameter NMR tubes prior to analysis and proceed to Step 4.

(C) Faecal water sample extraction and preparation for NMRTIMING ~ 2 days for a batch of 96 samples

(i) Defrost the stool aliquot

(ii) Weigh a representative amount (~15 g) of the homogenized stool sample

(iii) mix with a 2:1 ratio of water - net weight of faecal sample (µl − mg).

(iv) Vortex the mixture for 5 min and centrifuged at 4 °C at 18 000 g for 10 min.

(v) Store at ≤-20 °C until analysis.

(vi) Immediately prior to analysis, defrost at room temperature for 1 h and vortex the samples for 10 sec and centrifuge at 4 °C for 10 min at 18,000 *g*.

(vii) Transfer a total of 400 µl of supernatant into a new microcentrifuge tube containing 250 µl of the phosphate buffer (see REAGENT SET UP).

(viii) Vortex the mixture and centrifuge for 10 sec.

(ix) Pipette 600 µl of the mixture into 5mm NMR tubes and proceed to Step 4.

(D) Tissue, breastmilk and food sample preparation for NMR (extraction of polar and lipophilic metabolites). TIMING ~ 1-1.5 days per 96-well plate

(i) Defrost samples at room temperature and properly homogenize.

For breast milk samples, vortex vigorously to resuspend the milk fat globules.

For food samples, vortex vigorously to assure homogeneity of the sample.

CRITICAL STEP The use of electric homogenizer for homogenization of frozen tissue samples in solvents is recommended over the use of manual grinding method with a mortar because it provides less inter-sample variability. According to published literature,[13] the typical sample size is 100 mg (wet mass) although as little as 20 mg can be used. As

previously described for for breast milk[44] and tissue samples[13], the Folch extraction method requires:

(ii) Prepare ice-cold solvents: 400 ml of methanol, 200 ml of chloroform, 100 ml of water (amounts calculated for 100 samples)

(iii) For tissue samples, weigh intact frozen tissue (200 mg), then transfer into a glass vial.

(iv) Homogenize after adding 4 ml of methanol per gram of tissue and 0.85 ml g^{-1} water to the sample. Vortex the sample, then add 2 ml g^{-1} chloroform to the sample and vortex again.

(v) Add 2 ml g^{-1} chloroform and 2 ml g^{-1} water to the sample and vortex again.

(vi) Leave sample on ice or in the fridge for 15 min. Centrifuge at 1,000g for 15 min at 4°C. The solutions should now separate into an upper methanol/water phase (with polar metabolites) and a lower chloroform phase (with lipophilic compounds), separated by protein and cellular debris. Centrifuge again if still no clear separation.

(vii) In turn, transfer the lower layer first followed by the upper layer of each sample into separate glass vials. Remove the solvents from the samples using a speed vacuum concentrator or under a stream of nitrogen and proceed to Step 3.

E) SPE of urine samples . TIMING ~ 20 min per sample.

(i) Centrifuge urine sample for 10 min at 10,000g to remove particulates (4 °C).

(ii) Follow manufacturer's instructions to condition and equilibrate sorbent.

(iii) Acidify sample according to manufacturer's instructions.

(iv) Load sample onto sorbent.

(v) Wash according to manufacturer's instructions.

(vi) Elute according to the manufacturer's instructions.

(vii) Evaporate both elution samples to dryness.

(viii) Reconstitute sample in water.

(ix) Prepare samples into 5mm NMR tubes or glass LC vials according to the analytical platform to be used[85] and proceed to Step 4.

TROUBLESHOOTING

(F) LC-fraction collection of urine samples TIMING ~ 90 min per sample.

(i) Follow manufacturer's instructions to condition the LC and the column.

(ii) Centrifuge urine sample for 10 min at 10,000g to remove particulates (4 °C).

(iii) Lyophilize 5ml of urine sample and reconstitute in 500 µl of the original urine sample.

(iv) Vortex, sonicate and centrifuge the sample for 20 min at 16 000 × g

(v) Set up the LC method and the starting time, waiting time and frequency for fraction collection.

(vi) Inject the supernatant onto a HPLC column (reversed phase or similar according to the nature of the analyte to be separated) (7 × 2 µl) repeatedly

(vii) Collect the fractions in individual glass tubes.

(viii) Dry each fraction under a stream of nitrogen.

(ix) Reconstitute the dried fraction in 600 µl of H_2O water.

(x) For NMR analysis, mix 540 µl of the reconstituted fraction with 60 µl of NMR buffer (see REAGENT SET-UP).

(xi) For LC-MS analysis, inject 50 µl of the reconstituted fraction into the LC-MS system[11].

(xii) If several NMR-visible peaks that belong to other compounds are present in the same fraction, go back to step (v) and change the gradient in the LC method and follow the procedure. Alternatively, you can go back to step (i), install and condition a new LC column and follow the procedure.

TROUBLESHOOTING

OPTIONAL Additional processing of tissue, breastmilk and food extracts

3| Process the tissue, breastmilk and food extracts for NMR acquisition using option A for aqueous extracts/water-soluble metabolites or option B for lipophilic extracts/lipid metabolites.

(A) Preparation of aqueous extracts/water-soluble metabolites for NMR spectroscopy. TIMING ~ 1-1.5 h per 96-well plate

(i) Reconstitute the polar tissue extracts in either 580 µl of NMR buffer for tissue or breast milk (see REAGENT SET UP) or in D_2O containing TSP as a chemical shift reference (δ = 0).

(ii)Vortex samples and then centrifuge at 12,000g for 5 min.

(iii) Transfer 550 µl of the supernatant into an NMR tube prior to analysis and proceed to Step 4.

(B) Preparation of lipophilic extracts/lipid metabolites for NMR spectroscopy. TIMING ~ 1-1.5 h per 96-well plate

(i) Resuspend the lipophilic extracts in 580 ml deuterated NMR solvent (2:1 mixture of chloroform-d ($CDCl_3$) containing 0.03 vol/vol TMS, and CD_3OD) and then vortex. This method works well when running NMR experiments manually. For automated runs, we have found that resuspending the lipophilic tissue extracts in 580 ml deuterated $CDCl_3$ containing 0.03 vol/vol TMS only is robust and reliable with regard to locking and shimming.

(ii) After centrifugation (1,000g, 5 min), transfer 550 ml of the supernatant into an NMR tube and proceed to Step 4.

Data acquisition

4| Acquire NMR data using option A, and LC-MS data using option B. Set the temperature to 37°C, 310 K

(A) NMR Data acquisition. TIMING ~ 15 min per sample (from (i) to (vi)) and from 2 h to 2-3 days (from (i) to (vii))

(i) Change the SampleJet mode to the 5 mm shuttle automation mode.

(ii) ConFig. SampleJet preheating time to 1min at the appropriate temperature as previously described in NMR set up.

(iii) Load sample into preheating station.

(iv) Load sample into probe and allow the temperature to equilibrate.

(v) Tune and match the probe using the automation routine that involves: to set the RF carrier frequency offset value to the H_2O resonance, to determine the water saturation power, to determine the 90^0 pulse length at a given power level; to re-adjust the frequency offset for water signal suppression if necessary and to update the experiments with the optimised parameters before submitting the run.

(vi) Select suitable experimental pulse sequence for the samples as previously described in NMR SETUP: standard one-dimensional pulse sequence (noesygppr1d) for urine, breast milk, food, tissue and faecal samples; standard one-dimensional pulse sequence (noesygppr1d) and CPMG-presat (cpmgpr1d) for plasma and serum.

(vii) For any type of samples, select when appropriate *J*-resolved, and diffusion-edited.The 1D spectra are generally processed by applying a line broadening of 0.3–1 Hz and zero-filling by a factor of 2 to give 64k frequency domain data points.

TROUBLESHOOTING

(B) LC-MS DATA ACQUISITION TIMING ~2 h for a 96-well plate

(i) Follow manufacturer's instructions to condition the LC and the column.

(ii) Centrifuge 96-well plate or vials at 10,000g for 5 min (4 °C).

(ii) Load 96-well plate or vials into autosampler and maintain at 4 °C.

(iii) Select ESI ionization mode- positive or negative.

(iv) Carry out the accurate mass set up by infusing appropriate concentration of leucine enkephalin (or alternative lockmass solution) into instrument.

(v) Carry out the calibration setup by infusing sodium formate solution or alternative calibration solution into the instrument.

(vi) Follow setup procedures: Ion counts must be below 200 counts per second in continuum mode for the accurate mass and calibration setup (Capillary voltage and cone voltage will be adjusted until criteria are filled manually or automatically according to manufacturer options). A measure of the 'fit' of the calibration line to the experimental data is given in the error of the residual. As a general rule, the residual (in mDa) for each individual calibration point should be < 0.5 mDa.

(vii) Select suitable gradient according to the sample and the method.

TROUBLESHOOTING

TIMING

Workflows 1 and 2

These workflows are based on computational/statistical calculations that require only a few seconds depending on computer power.

Workflows 3 and 4

Procedure step 1 (A, B), sample preparation for urine, serum or plasma takes 1.5–2.5 h per 96-well plate. Sample preparation can be automated and it takes 1–3 h[27] to prepare a batch of 96 samples depending on the type of preparation robot.

Procedure step 1(C), sample preparation from frozen faecal sample it requires 2 days to prepare a batch of 96 samples (including the weight of the sample, the extraction and the adition of the buffer)[80].

Procedure step 1 (D), food and tissue samples require a longer time due to the sample homogenization, solvent extraction process and solvent evaporation, and take 1-1.5 days per 96 well plate[8,41].

Procedure step 3 (A), for the NMR experiments, a throughput of around 78 serum/plasma samples per day can be achieved if a complete set of three experiments such as 1D ^{1}H-NMR, 1D CPMG, 2D *J*-RES (NOESY_presat pulse sequence, J-resolved sequence, Carr-Purcell-Meiboom-Gill (CPMG) spin-echo sequence)[27] is applied. 150 urine, faecal, food and tissue samples can be measured per day when performing a set of two experiments (NOESY_presat pulse sequence and *J*-resolved sequence)[27]. For the 2D-NMR experiments, the time varies according to the type of experiment: 2D long^{1}H-^{1}H *J*-RES (2 h per sample), ^{1}H-^{13}C HSQC (2-3 days per sample), ^{1}H-^{13}C HMBC (2-3 days per sample), ^{1}H-^{1}H COSY (10-12 h per sample), ^{1}H-^{1}H TOCSY (1 day per sample) (Table 1) .

Workflow 5

Procedure step 2 (E), It requires approximately 20 min per sample to perform SPE experiments depending on the type of cartridge and its requirements.

Liquid-liquid extraction requires 15-30 min per sample but depends on the type of solvents use. For example, the Folch procedure[86] takes 15-20 min per sample.

Freeze drying experiments take 12-24 h per sample depending on the amount of sample to be dried.

Procedure step 2 (E,F),It requires approximately 2-10 days, depending on the abundance of the metabolite in the sample to ensure enough quantitative material to acquire good NMR experiments.

Workflow 6

It requires approximately 1h per sample including sample preparation

Workflow 7

It requires approximately 2-3h per sample including sample preparation

Workflow 8

GC-MS metabolic profiling analysis requires 12-24 hours for sample derivatization and 30-40 min for sample analysis.

CE-MS metabolic profiling analysis requires 5 min for sample preparation and 20-25 min for sample analysis.

TROUBLESHOOTING

Troubleshooting advice is summarised in **Table 2**.

TABLE 1 | Summary of 1D and 2D NMR experiments for acquiring metabolic profiling data and metabolite identification.

Experiment	***Aim of the experiment***	***Timing***	***Notes***
1D ^{1}H-NMR (NOESY_presat pulse sequence)	To acquire metabolic profiles	***As part of the routine for metabolic profiling analysis:*** • 96 amples can be measured within 24 h, when running a NOESY-presat and JRes (~15 min per sample). • 72 samples can be measured within 24 h when running a NOESYpresat, CPMG, and JRES (~19 min per sample). • Longer acquisition of 2D ^{1}H-^{1}H *J*-RES will require around 2 h per sample.	The first experiment to apply in metabolic profiling analysis to urine, blood, faeces, breast milk and tissue samples.

1D CPMG (Carr-Purcell-Meiboom-Gill (CPMG) spin-echo sequence)	To acquire metabolic profiles	As above	At times, it might be appropriate to use the CPMG-presat sequence for water-soluble metabolites to attenuate the NMR signals of any remaining proteins.
$^{1}H-^{1}H$ *J*-resolved (JRes)	To assess the values of J couplings (Hz), and the multiplicity of the signals by tackling the problem of feature overlap.	As above	If possible, it is recommended to acquire 2D J-Res spectra just after the 1D ^{1}H-NMR or 1D CPMG experiments in order to be able to perform RED-STORM.
$^{1}H-^{1}H$ COrrelation SpectroscopY (COSY)	With this experiment is possible to see short range correlations ($^{2-3}J_{H-H}$) by measuring of spin-spin coupling up to 4 bonds away, but mainly detects neighbouring protons.	It depends on the sample concentration but requires ~10-12 h per sample.	
$^{1}H-^{1}H$ TOtal Correlation SpectroscopY (TOCSY)	With this experiment is possible to see the homonuclear long-range correlations. ($^{2-n}J_{H-H}$) by measuring spin-spin coupling of up to 6 or 7 bonds distance depending on the length of the spin-lock time in the pulse sequence and displays proton-proton couplings on the off-diagonal allowing the connections between adjacent protons to be reconstructed	It depends on the sample concentration but requires ~24 h per sample	.
$^{1}H-^{13}C$ Heteronuclear Single-Quantum Coherence (HSQC)	With this experiment is possible to see direct coupling between protons and a second nucleus with spin I=1/2, usually ^{13}C.	It depends on the sample concentration but requires ~2-3 days per sample.	
$^{1}H-^{13}C$ Heteronuclear Multiple Bond-Correlation (HMBC)	Heteronuclear long-range correlations. The HMBC pulse sequence detects $^{1}H-^{13}C$ connections across multiple, typically up to	It depends on the sample concentration but requires ~1day per sample.	

	four, bonds. With this experiment is possible to see quaternary carbons including carbonyl and ipso carbons.		

TABLE 2 | Troubleshooting table

Procedure Step	Problem	Possible reason	Solution
Serum sample collection, step 1B	Incompletely clo tted sample after spinning the vacutainer tubes for 15 minutes.	Sample requires longer time to clot	Spin the vacutainer tubes up to clotting time at 4°C
Breast milk sample preparation,step 1D	The defrosted breast milk sample is split in two layers	Lack of sample homogeneity	Vortexe vigorously to resuspend the milk fat globules
NMR data acquisition, step 4A	Automatic locking failure	Buffer is missing	Repeat sample preparation, mixing the right sample volume with the buffer
	Baseline rolling ('wiggles') after pre-concentrating the sample	Sample concentration has increased too much in relation to the original sample	Select automatic receiver gain adjustment
	Water presaturation failure	Sample concentration is too low	Pre-concentrate the sample or recalibrate water suppression
	Peak shift and/or shape is similar but not identical to those suggested in data tables or does not precisely overlay with a spectrum of the standard compound	Changes in pH can impact on peak position and pH can alter over time even when buffered due to precipitation of compounds	Readust pH and rerun and/or spike in a low concentration of the outative metabolite
SPE data	Steady stream	Sample was loaded	Load proper

acquisition, step 4B	when loading the sample	too fast	flow rate (follow manufacture instructions)
	Poor Recovery	Too much sample was loaded on the cartridge size Too strong wash step Too weak elution Cartridge dried out	Reduce sample load or use more sorbent Optimize wash step (follow manufacture instructions) Optimize elution step (follow manufacture instructions) Use a new cartridge and start loading the sample when there is still a thin layer of water above the sorbent
	Sorbent was overloaded	High content of salts or organic matter in the matrix	Reduce sample load or use more sorbent
LC-NMR-MS data acquisition, step 4C	Poor MS resolution	Sample was overloaded into the MS systems	Inject smaller amount of sample in the mass spectrometer. Use an efficient postcolumn splitter to divert a small portion of the flow into the mass spectrometer if you are using an inline LC-NMR-MS system

LC-MS data acquisition, step 4D	High Backpressure	Particulate matter from sample may have caused a blockage in: a) column b)capillary or tubing of MS source c) capillary tubing of LC d) LC injection system	Replace column Replace MS source tubing or capillary Replace LC tubing Replace LC injection tubing
	Few or absence of peaks	Failed injection Needle blockage Sample concentration too low	Flush needle Reinject sample Replace needle Repeat sample preparation Concentrate sample
	Poor chromatographic peak shape	Column contamination or degradation Overloading of sample	Clean or replace column Dilute sample or improve sample preparation
	Drop in baseline	Ion suppression, perhaps due to high salt levels in sample	Improve sample preparation with salt removal Optimize chromatographic gradient to minimize coelution of peaks if possible
	Carry-over	Selected wash solvents may not be appropriated Too high injection volume Lack of chromatography	Choose suitable wash solvents Optimize injection volume and concentration of components in the sample

		optimization	Optimize chromatographic gradient
	Loss of sensitivity	Matrix suppression Poor recovery	Improve sample preparation with salt-removal
	Unsteady beam	Capillary/sample cone voltages not optimal Capillary is protruding too far from end of probe Probe is too far into source Liquid chromatography (LC) solvent flow is not correct/steady Solvents have not been adequately degassed Desolvation/nebulizer gas flow is not steady Desolvation temperature is not set correctly for liquid flow rate used	Tune sample cone and capillary Change length of capillary protruding from probe Move probe away from source Degas solvent, reset and remeasure the flow rate Check and adjust nitrogen supply pressure Check manual for guidelines Check and adjust desolvation temperature Check manual for guidelines
	Loss of sensitivity	Ion source is dirty Matrix suppression Component failure	Clean the source according to manufacturer guidelines Improve sample preparation such as salt-removal.

		of MS system	Arrange engineer visit
	High chemical or electronic noise levels	Signal threshold set too low Detector damaged and producing micro discharges	Reduce detector voltage Arrange engineer visit

ANTICIPATED RESULTS

We present here a system for NMR-based identification of metabolites in biofluids and extracted biological samples based on various statistical and analytical resources .

An exemplification of the use of the workflows presented in Fig.1 is given for the identification of various metabolites. For example, the identification of ascorbic acid (Fig.2a-g) in urine ^{1}H-NMR spectra from participants who were characterised by high BMI [10] was achived after combining workflows 1,2 and 4. Unconclusive results were obtained after performing STOCSY (workflow 1, Fig.2a-c) and therefore STORM (workflow 2, Fig.2d) was conducted on a subset of only 33 out of 1,880 spectra to reveal a correlation between the unknown signal and two multiplet signals δ 4.03 (ddd) of chiral CH and δ 3.74 (m) of CHH'OH], possibly belonging to the same molecule tentatively identified as ascorbic acid. This assignment was subsequently confirmed using a collection of 2D NMR experiments (Workflow 4) to measure the sample containing the highest signal intensity such as: 2D-Jres (Fig. 2e) that confirmed the multiplicity of the signals at δ 4.03 (ddd) and 3.74 (m), ^{1}H−^{1}H TOCSY (Fig. 2f) that showed structural correlation between δ 4.03, 3.74, and 4.52 (d) and an ^{1}H−^{13}C HSQC pulse sequence (Fig. 2g) applied on the urine sample and the ascorbic standard. Finally, structural elucidation of the compound was confirmed by spiking of the urine samples with higher volumes of the same ascorbic acid pure solution (Fig. 2h).

A different approach was taken based on the workflows 1,2,4 and 5 presented in Fig. 1 to successfully identify of *N*^{1}-Methyl-2-pyridone-5-carboxamide (2PY) and *N*^{1}-Methyl-4-pyridone-5-carboxamide (4 PY) in urine of C57BL/6 mice (Fig. 3a-h). Urinary 1H NMR

spectrum were acquired and modelled using OPLS-DA. The unknown NMR peaks at 6.66, 6.69, 7.83, 7.97 and 8.32 ppm were found positively correlated with High Fatfed mice in the OPLS-DA loadings plot (Fig. 3a). The signal at 6.66 ppm was selected as a driver to conduct STOCSY (workflow 1) analysis which showed several peaks significantly correlated with the peak at 6.66 ppm (Fig. 3b). However, the application of 2D-NMR experiments (Workflow 4) showed incomplete assignment of the unknown or potential unknowns with cross-correlation peaks at δ= 6.66-7.97 ppm and δ=6.69-7.83 ppm (Fig. 3c). Therefore, Solid Phase Extraction (workflow 5) was performed to isolate the unknown signals. As a result, the signals of the unknown were isolated and observed on the spectral overlay of the pooled urine after elution (25% MeOH 75% HCl 0.1 M) on C18 cartridge (in red) and the standard 1D ^{1}H NMR spectrum of intact pool urine (in blue) (Fig. 3d) . A catalogue of 2D-NMR experiments were performed on the first elution that contains the isolated signals. Cross correlation peaks at δ=6.66-7.97 ^{1}H-^{1}H ppm, δ= 7.97-8.32 ^{1}H-^{1}H ppm, δ= 6.69-7.83 ^{1}H-^{1}H ppm, δ= 7.83-8.54 ^{1}H-^{1}H ppm were observed on the zoom of 2D ^{1}H-^{1}H COSY spectrum (Fig. 3e) between 6 and 9 ppm of the C18 column fraction of pooled urine. Moreover, cross correlation peaks at δ= 6.66-120.7 ppm, δ=6.69-123 ^{1}H-^{13}C ppm, 7.83-130 ^{1}H-^{13}C ppm, δ= 7.97-142 ^{1}H-^{13}C ppm, δ=8.32-144.5 ^{1}H-^{13}C ppm, δ= 8.54-149.29 ^{1}H-^{13}C ppm were observed on the zoom of 2D ^{13}C-^{1}H HSQC spectrum between δ= 6 and 9 ppm of the C18 column fraction of pooled urine (Fig. 3f). Finally, standards were spikied by adding repeatedly the same volume (2 µl) of pure 4PY solution at 2mM and 3mM in the C18 column fraction of pooled urine (Fig. 3g) and (2 µl) of pure 2PY solution at 2mM and 3mM in the C18 column fraction of pooled urine (Fig. 3h).

Finally, another strategy was taken for the identification of the dietary biomarker of onion intake [10] (Fig. 4a-c, Fig. 5a-e). Samples were acquired by 1H NMR analysis and the spectrum showed a characteristic multiplet associated with onion intake during an in-patient crossed-over controlled clinical trial (Fig. 4a). STORM (workflow 2, Fig.4b) was conducted and revealed 3 spectral peaks related to the same multiplet. As a result of performing RED-STORM (workflow 3, Fig. 4c) two more multiplets were identified in the J-resolved pseudospectrum. LC-NMR-MS (workflow 6) was then conducted on the urine sample with the highest amount of the unknown that needed to be lyophilized overnight and reconstituted in 500 µl of the original urine sample and then injected in the LC column. Every 29 s one fraction was collected, dried and analysed by 1H-NMR. As a result, the unknown metabolite was isolated in one of the fractions (Fig. 5a). 50 µl of the fraction containing the unknown metabolite was also analysed by reverse-phase LC-MS. The LC-MS (positive mode) chromatogram (Fig. 5b) and the corresponding total ion chromatogram of the fraction (Fig. 5**c**) revealed $C_8H_{13}NO_4S$ as likely elemental composition. Furthermore, 2D-NMR analysis (workflow 4) was performed on the fraction containing the unknown which allowed to finally elucidate the NMR signals and chemical shifts (Fig. 5d) and the “unknown unknown” was identified as *N*-acetyl-S—1-*Z*-propenyl-cysteine-sulfoxide (NAcSPCSO) (Fig.5e).

We anticipate that routine implementation of this system in metabolic profiling studies will enhance biological interpretation of pre-clinical, clinical, epidemiological and nutritional studies and will allow broad coverage of the metabolome and consequently provide deeper insight into mechanistic pathways underlying disease aetiology. The adoption of improved metabolite identification workflows will result in increased population of spectral standards databases, which will enable researchers within the field to make more comprehensive assingments in a shorter timeframe. Improved annotation and correlation of accurately identified molecules associated with specific physiological or pathological conditions will also improve the generation of reliable biomarker panels for human, animal and plant diseases.

ACKNOWLEDGMENTS. I.G.-P. is supported by a National Institute for Health Research (NIHR) fellowship (NIHR-CDF-2017-10-032). J.M.P. is supported by a Rutherford Fund Fellowship at Health Data Research (HDR) UK (MR/S004033/1). G.F. is an NIHR Senior Investigators. P.E. is Director of the Medical Research Council (MRC) Centre for Environment and Health (MR/L01341X/1), and acknowledges support from the NIHR Imperial Biomedical Research Centre, and the NIHR Health Protection Research Unit in Health Impact of Environmental Hazards (HPRU-2012-10141). P.E. is supported by the UK Dementia Research Institute supported by UK DRI Ltd which is funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK. INTERMAP is supported by the U.S. National Heart, Lung and Blood Institute (grants R01-HL050490 and R01-HL084228), the Chicago Health Research Foundation, and national agencies in Japan (grant [A] 090357003), People's Republic of China, and the United Kingdom (R2019EPH). Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre (BRC) and the UK MEDical BIOinformatics partnership (MR/L01632X/1).

AUTHOR CONTRIBUTIONS.

The protocol was written by I.G.-P., J.M.P., J.L., I.S.C., G.F., P.E., E.H. and J.K.N.

CONFLICT OF INTEREST. The authors declare no conflict of interest. Funders had no role in study design.

Fig. 1 | Overview of our system for metabolite identification based on a combination of analytical and statistal workflows.

This protocol describes eight different workflows to be followed in a recommended sequential order according to level of difficulty to identify molecular species from NMR-based metabolic phenotyping studies. This multiplatform strategy is based on statistical spectroscopic tools, 2D-NMR spectroscopic analysis, separation and pre-concentration techniques, multiple analytical platforms and existing databases.

a, Summary of the overall metabolic profiling workflow that comprises sample preparation, 1H-NMR data acquisition, data modelling and assessment of biomarkers to be structurally elucidated.

b, The standard approach (workflow 1) uses STOCSY followed by database matching and spiking of commercial standards.

c, Alternative approaches (workflows 2-8) can be followed if the standard approach fails or the method is not appropriate (e.g. in the cases which no correlations were found, there is a degree of overlap of signals, the unknown is in low concentration or the information provided is insufficient). A Dotted line indicates the possibility of altering the recommended sequential order of the workflows according to accessibility to data previously acquired and cost and time efficiency criteria.The first two variants require existing data from different types of statistical spectroscopy – STORM (workflow 2) and RED-STORM (workflow 3). This can be followed by applying traditional 2D NMR pulse sequences (workflow 4) to extract more information about molecular structure, perform physical concentration or separation of the 'unknown' molecule before re-analysis using lyophilisation, SPE, liquid-liquid extraction (workflow 5), LC-NMR-MS (workflow 6) or LC-SPE-NMR-MS (workflow 7).

Positive identification of the unknown metabolite should be performed by spiking authentic chemical standards into the biosample. In the case of identifying an 'unknown unknown' compound, the name, structure, CAS number and or International Chemical Identifier should be deposited in a database. If the standards are not commercially available, or it is not feasible to synthesis the compound, it will remain as a putatively annotated compound or putatively identified within a compound class.

Fig. 2 | Data from the metabolome-wide assocaton (MWA) study of human urine samples (analysed by 1D NMR, STOCSY, STORM, 2D NMR and spiking of standard**) from the INTERMAP epidemiological study.**

a, **Unknown metabolites linked to BMI**: Expanded region δ 4.05 to 4.00 from the median 600 MHz ^{1}H-NMR spectrum of urine from 1880 non-diabetic U.S. INTERMAP participants, shows in green significant positive associations with BMI. **b**-**c, Workflow 1**: Unclear results obtained after performing STOCSY of the most significant signals δ 4.04 and 4.02. **d**, **Workflow 2**: STORM revealed a correlation between the unknown signal and two multiplet signals δ 4.03 (ddd) of chiral CH and δ 3.74 (m) of CHH'OH], possibly belonging to the same molecule. **e-g**, **Workflow 4**: Results from a series of 2D-NMR expermients performed on the

urine sample with the highest amount of unknown metabolite. This result contributed to the structural elucidation of the unknown metabolite by enhancing the information acquired in Workflow 2. **e**, **Workflow 4**: Untilted JRES spectrum of the region from δ 4.75 to δ 3.63, shows the multiplicity of the signals at δ 4.03 (ddd) and 3.74 (m). **f**, **Workflow 4**: $^{1}H-^{1}H$ TOCSY spectrum (δ 4.75 to 3.63) shows structural correlation between δ 4.03, 3.74, and 4.52 (d). **g**, **Workflow 4:** $^{1}H-^{13}C$ HSQC spectra (F1, δ 87.5 to 32.5; F2, δ 4.75 to 3.63) of (red) urine sample overlapping with the $^{1}H-^{13}C$ HSQC spectra of (blue) ascorbic acid standard. **h**, **Spiking of standard:** Structural elucidation of the compound is confirmed by spiking of the urine samples with higher volumes of the same ascorbic acid pure solution. The figure shows a zoomed region of the ^{1}H-NMR spectra of spike-in experiments of ascorbic acid in the urine sample with the multiplicity shown underneath. The urine sample spectrum is shown in red, and spectra with added 0.95 and 1.42 µmol of 1mmol ascorbic acid are shown in green and blue, respectively. Key: d, doublet; ddd, doublet of doublets of doublets; m, multiplet.

The human data shown in this figure belong to a study approved by the London‐Brent Research Ethics Committee and performed in accordance with the Declaration of Helsinki (13/LO/0078). Figure adapted from Posma et al. [10]

Fig. 3 | Strategies for the identification of *N*^{1}-Methyl-2-pyridone-5-carboxamide (2 PY) and *N*^{1}-Methyl-4-pyridone-5-carboxamide (4 PY) in urine of C57BL/6 mice. Flow diagram outlining the process applied (analysed by 1D NMR, STOCSY, 2D NMR, SPE, 2D NMR and spiking of standard**) from based on Fig.1 to conduct the metabolite identification**

a, OPLS-DA loadings plot indicating the unknown NMR peaks at 6.66, 6.69, 7.83, 7.97 and 8.32 ppm positively correlated with High Fatfed mice.

b, Workflow 1: STOCSY analysis shows several peaks significantly correlated with the driving peak at 6.66 ppm. **c**, Workflow 4: the application of 2D-NMR experiments shows incomplete assignment of the unknown or unknowns. Zoom of 2D ^{1}H-^{1}H COSY spectrum between δ= 6 and 9 ppm of pooled urine sample showing cross-correlation peaks at δ= 6.66-7.97 ppm and δ=6.69-7.83 ppm. **d**, Workflow 5: Solid Phase Extraction (SPE) is performed to isolate the unknown signals. The Spectral overlay of standard 1D ^{1}H NMR spectrum of intact pool urine (in blue) and pooled urine after elution (25% MeOH 75% HCl 0.1 M) on C18 cartridge (in red) showing the isolated signals. **e-f**, Workflow 4: a catalogue of 2D-NMR experiments are performed on the first elution that contains the isolated signals. **e**, Zoom of 2D ^{1}H-^{1}H COSY spectrum between 6 and 9 ppm of the C18 column fraction of pooled urine shows cross correlation peaks at δ=6.66-7.97 ^{1}H-^{1}H ppm, δ= 7.97-8.32 ^{1}H-^{1}H

ppm, δ= 6.69-7.83 ^{1}H-^{1}H ppm, δ= 7.83-8.54 ^{1}H-^{1}H ppm. **f**, Zoom of 2D ^{13}C-^{1}H HSQC spectrum between δ= 6 and 9 ppm of the C18 column fraction of pooled urine shows cross correlation peaks at δ= 6.66-120.7 ppm, δ=6.69-123 ^{1}H-^{13}C ppm, 7.83-130 ^{1}H-^{13}C ppm, δ= 7.97-142 ^{1}H-^{13}C ppm, δ=8.32-144.5 ^{1}H-^{13}C ppm, δ= 8.54-149.29 ^{1}H-^{13}C ppm. **g-h**, Spiking of the standards by adding repeatedly the same volume (2 µl) of two standard solutions of different concentrations. **g**, Spiking of pure 4PY solution at 2mM and 3mM in the C18 column fraction of pooled urine. **h**, Spiking of pure 2PY solution at 2mM and 3mM in the C18 column fraction of pooled urine.

The animal data show in this figure belongs to a study conducted according to local animal welfare policy and approved by Swiss governmental veterinary offices (authorization number VD-2231)

Fig. 4 | Statistical Spectroscopic strategies for the identification of an "unknown unknown" molecule in human urine: N-acetyl-S-(1Z)-propenyl-cysteine-sulfoxide (NAcSPCSO), biomarker of onion intake a,

Flow diagram outlining the process applied (analysed by 1D NMR, STORM, RED-STORM, LC-NMR-MS, 2D NMR and spiking of standard**) based on Fig.1 to conduct the metabolite identification**. **a,** Section of 600 MHz 1H-NMR spectrum shows a characteristic multiplet associated with onion intake of one volunteer during an in-patient crossed-over controlled clinical trial. **b**, Workflow 2: STORM analysis shows 3 spectral peaks related to tentative multiplet. **c**, Workflow 3: application of RED-STORM identifies two more multiplets compared to STORM that are highlighted in red in the J-resolved pseudospectrum. The human data show in this figure belongs to a study approved by the London‐Brent Research Ethics Committee and performed in accordance with the Declaration of Helsinki (13/LO/0078). Adapted from Posma et al. [10].

Fig. 5 | LC-NMR-MS strategies for the identification of an "unknown unknown" molecule in human urine: N-acetyl-S-(1Z)-propenyl-cysteine-sulfoxide (NAcSPCSO), biomarker of onion intake , Workflow 6: LC-NMR-MS, the urine sample with the highest amount of the unknown is lyophilized overnight and reconstituted in 500 µl of the original urine sample and then injected in the LC column. Every 29 s one fraction is collected, dried

and prepared to be analysed by 1H-NMR. **a**, 1H NMR spectrum of the fraction with the highest concentration of the unknown metabolite, with multiplets assigned and integrals calculated. **b,** Total ion chromatogram of the fraction containing the unknown metabolite after LC-MS analysis. **c,** The LC-MS (positive mode) chromatogram showing the $C_8H_{13}NO_4S$ as likely elemental composition. **d**, Workflow 4: 2D-NMR analysis performed on the fraction containing the unknown. The table shows the identification of the NMR signals and chemical shifts of the unknown. **e**, Chemical structure of the "unknown unknown": *N*-acetyl-S--1-*Z*-propenyl-cysteine-sulfoxide (NAcSPCSO). The human data show in this figure belongs to a study approved by the London‐Brent Research Ethics Committee and done in accordance with the Declaration of Helsinki (13/LO/0078). Adapted from Posma et al. [10].

REFERENCES

1 Holmes, E., Wilson, I. D. & Nicholson, J. K. Metabolic phenotyping in health and disease. *Cell* **134**, 714-717, doi:10.1016/j.cell.2008.08.026 (2008).

2 Nicholson, J. K., Lindon, J. C. & Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica; the fate of foreign compounds in biological systems* **29**, 1181-1189, doi:10.1080/004982599238047 (1999).
3 Fiehn, O. Metabolomics - the link between genotypes and phenotypes. *Plant Mol Biol* **48**, 155-171, doi:Doi 10.1023/A:1013713905833 (2002).
4 Nicholson, J. K. & Wilson, I. D. High-Resolution Proton Magnetic-Resonance Spectroscopy of Biological-Fluids. *Progress in nuclear magnetic resonance spectroscopy* **21**, 449-501, doi:Doi 10.1016/0079-6565(89)80008-1 (1989).
5 Nicholson, J. K. *et al.* Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *The Biochemical journal* **217**, 365-375 (1984).
6 Bales, J. R., Higham, D. P., Howe, I., Nicholson, J. K. & Sadler, P. J. Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. *Clinical chemistry* **30**, 426-432 (1984).
7 Wilson, I. D., Wade, K. E. & Nicholson, J. K. Analysis of Biological-Fluids by High-Field Nuclear Magnetic-Resonance Spectroscopy. *Trac-Trend Anal Chem* **8**, 368-374, doi:Doi 10.1016/0165-9936(89)85075-7 (1989).
8 Belton, P. S. *et al.* Use of high-field H-1 NMR spectroscopy for the analysis of liquid foods. *Journal of agricultural and food chemistry* **44**, 1483-1487, doi:Doi 10.1021/Jf950640z (1996).
9 Cloarec, O. *et al.* Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Analytical chemistry* **77**, 1282-1289, doi:10.1021/ac048630x (2005).
10 Posma, J. M. *et al.* Subset optimization by reference matching (STORM): an optimized statistical approach for recovery of metabolic biomarker structural information from 1H NMR spectra of biofluids. *Analytical chemistry* **84**, 10694-10701, doi:10.1021/ac302360v (2012).
11 Posma, J. M. *et al.* Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. *Analytical chemistry* **89**, 3300-3309, doi:10.1021/acs.analchem.6b03324 (2017).
12 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. *Analytical chemistry* **67**, 793-811, doi:Doi 10.1021/Ac00101a004 (1995).
13 Beckonert, O. *et al.* Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* **2**, 2692-2703, doi:10.1038/nprot.2007.376 (2007).
14 Dona, A. C. *et al.* A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Computational and structural biotechnology journal* **14**, 135-153, doi:10.1016/j.csbj.2016.02.005 (2016).
15 Godejohann, M., Tseng, L. H., Braumann, U., Fuchser, J. & Spraul, M. Characterization of a paracetamol metabolite using on-line LC-SPE-NMR-MS and a cryogenic NMR probe. *J Chromatogr A* **1058**, 191-196, doi:10.1016/j.chroma.2004.08.091 (2004).
16 Shockcor, J. P. *et al.* Combined HPLC, NMR spectroscopy, and ion-trap mass spectrometry with application to the detection and characterization of xenobiotic and endogenous metabolites in human urine. *Analytical chemistry* **68**, 4431-4435, doi:Doi 10.1021/Ac9606463 (1996).
17 Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S. & Zhang, Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic & biomolecular chemistry* **3**, 1832-1834, doi:10.1039/b502828k (2005).
18 Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics :*

Official journal of the Metabolomic Society **3**, 211-221, doi:10.1007/s11306-007-0082-2 (2007).
19 Wishart, D. S. Computational strategies for metabolite identification in metabolomics. *Bioanalysis* **1**, 1579-1596, doi:10.4155/bio.09.138 (2009).
20 Ellinger, J. J., Chylla, R. A., Ulrich, E. L. & Markley, J. L. Databases and Software for NMR-Based Metabolomics. *Current Metabolomics* **1**, doi:10.2174/2213235X11301010028 (2013).
21 Wishart, D. S. *et al.* HMDB: a knowledgebase for the human metabolome. *Nucleic acids research* **37**, D603-610, doi:10.1093/nar/gkn810 (2009).
22 Ulrich, E. L. *et al.* BioMagResBank. *Nucleic acids research* **36**, D402-408, doi:10.1093/nar/gkm957 (2008).
23 Akiyama, K. *et al.* PRIMe: a Web site that assembles tools for metabolomics and transcriptomics. *In silico biology* **8**, 339-345 (2008).
24 Wishart, D. S. Quantitative metabolomics using NMR. *Trac-Trend Anal Chem* **27**, 228-237, doi:10.1016/j.trac.2007.12.001 (2008).
25 Simpson, A. J., McNally, D. J. & Simpson, M. J. NMR spectroscopy in environmental research: from molecular interactions to global processes. *Progress in nuclear magnetic resonance spectroscopy* **58**, 97-175, doi:10.1016/j.pnmrs.2010.09.001 (2011).
26 Dalisay, D. S. & Molinski, T. F. NMR Quantitation of Natural Products at the Nanomole Scale. *J Nat Prod* **72**, 739-744, doi:10.1021/np900009b (2009).
27 Dona, A. C. *et al.* Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Analytical chemistry* **86**, 9887-9894, doi:10.1021/ac5025039 (2014).
28 Kumar, D. Nuclear Magnetic Resonance (NMR) Spectroscopy For Metabolic Profiling of Medicinal Plants and Their Products. *Crit Rev Anal Chem* **46**, 400-412, doi:10.1080/10408347.2015.1106932 (2016).
29 Fonville, J. M. *et al.* Evaluation of full-resolution J-resolved 1H NMR projections of biofluids for metabonomics information retrieval and biomarker identification. *Analytical chemistry* **82**, 1811-1821, doi:10.1021/ac902443k (2010).
30 Ludwig, C. & Viant, M. R. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical analysis : PCA* **21**, 22-32, doi:10.1002/pca.1186 (2010).
31 Viant, M. R. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochemical and biophysical research communications* **310**, 943-948 (2003).
32 Foxall, P. J. D., Parkinson, J. A., Sadler, I. H., Lindon, J. C. & Nicholson, J. K. Analysis of Biological-Fluids Using 600 Mhz Proton Nmr-Spectroscopy - Application of Homonuclear 2-Dimensional J-Resolved Spectroscopy to Urine and Blood-Plasma for Spectral Simplification and Assignment. *J Pharmaceut Biomed* **11**, 21-31, doi:Doi 10.1016/0731-7085(93)80145-Q (1993).
33 Liu, M., Nicholson, J. K. & Lindon, J. C. High-resolution diffusion and relaxation edited one- and two-dimensional 1H NMR spectroscopy of biological fluids. *Analytical chemistry* **68**, 3370-3376 (1996).
34 Spraul, M., Nicholson, J. K., Lynch, M. J. & Lindon, J. C. Application of the One-Dimensional Tocsy Pulse Sequence in 750 Mhz H-1-Nmr Spectroscopy for Assignment of Endogenous Metabolite Resonances in Biofluids. *J Pharmaceut Biomed* **12**, 613-618, doi:Doi 10.1016/0731-7085(93)E0028-L (1994).
35 Lindon, J. C., Nicholson, J. K. & Wilson, I. D. Directly coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development. *Journal of chromatography. B, Biomedical sciences and applications* **748**, 233-258 (2000).
36 Noda, I. Generalized 2-Dimensional Correlation Method Applicable to Infrared, Raman, and Other Types of Spectroscopy. *Appl Spectrosc* **47**, 1329-1336, doi:Doi 10.1366/0003702934067694 (1993).

37 Robinette, S. L., Lindon, J. C. & Nicholson, J. K. Statistical spectroscopic tools for biomarker discovery and systems medicine. *Analytical chemistry* **85**, 5297-5303, doi:10.1021/ac4007254 (2013).
38 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).
39 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).
40 Elliott, P. *et al.* Urinary metabolic signatures of human adiposity. *Science translational medicine* **7**, 285ra262, doi:10.1126/scitranslmed.aaa5680 (2015).
41 Garcia-Perez, I. *et al.* An Analytical Pipeline for Quantitative Characterization of Dietary Intake: Application To Assess Grape Intake. *Journal of agricultural and food chemistry* **64**, 2423-2431, doi:10.1021/acs.jafc.5b05878 (2016).
42 Garcia-Perez, I. *et al.* Bidirectional correlation of NMR and capillary electrophoresis fingerprints: a new approach to investigating Schistosoma mansoni infection in a mouse model. *Analytical chemistry* **82**, 203-210, doi:10.1021/ac901728w (2010).
43 Garcia-Perez, I. *et al.* Urinary metabolic phenotyping the slc26a6 (chloride-oxalate exchanger) null mouse model. *Journal of proteome research* **11**, 4425-4435, doi:10.1021/pr2012544 (2012).
44 Andreas, N. J. *et al.* Multiplatform characterization of dynamic changes in breast milk during lactation. *Electrophoresis* **36**, 2269-2285, doi:10.1002/elps.201500011 (2015).
45 Garcia-Perez, I. *et al.* Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. *Lancet Diabetes Endo* **5**, 184-195, doi:10.1016/S2213-8587(16)30419-3 (2017).
46 Posma, J. M. *et al.* Optimized Phenotypic Biomarker Discovery and Confounder Elimination via Covariate-Adjusted Projection to Latent Structures from Metabolic Spectroscopy Data. *Journal of proteome research* **17**, 1586-1595, doi:10.1021/acs.jproteome.7b00879 (2018).
47 Trygg, J., Holmes, E. & Lundstedt, T. Chemometrics in metabonomics. *Journal of proteome research* **6**, 469-479, doi:10.1021/pr060594q (2007).
48 Baranovicova, E. *et al.* NMR metabolomic study of blood plasma in ischemic and ischemically preconditioned rats: an increased level of ketone bodies and decreased content of glycolytic products 24 h after global cerebral ischemia. *Journal of physiology and biochemistry*, doi:10.1007/s13105-018-0632-2 (2018).
49 Scott, I. M. *et al.* Merits of random forests emerge in evaluation of chemometric classifiers by external validation. *Analytica chimica acta* **801**, 22-33, doi:10.1016/j.aca.2013.09.027 (2013).
50 Cavill, R. *et al.* Genetic algorithms for simultaneous variable and sample selection in metabonomics. *Bioinformatics* **25**, 112-118, doi:10.1093/bioinformatics/btn586 (2009).
51 Di Anibal, C. V., Callao, M. P. & Ruisanchez, I. 1H NMR variable selection approaches for classification. A case study: the determination of adulterated foodstuffs. *Talanta* **86**, 316-323, doi:10.1016/j.talanta.2011.09.019 (2011).
52 Wang, T. *et al.* Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC bioinformatics* **10**, 83, doi:10.1186/1471-2105-10-83 (2009).
53 Balabin, R. M., Safieva, R. Z. & Lomakina, E. I. Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. *Analytica chimica acta* **671**, 27-35, doi:10.1016/j.aca.2010.05.013 (2010).
54 Tiwari, P., Rosen, M. & Madabhushi, A. A hierarchical spectral clustering and nonlinear dimensionality reduction scheme for detection of prostate cancer from magnetic resonance spectroscopy (MRS). *Medical physics* **36**, 3927-3939, doi:10.1118/1.3180955 (2009).

55 Fotiou, M. *et al.* (1)H NMR-based metabolomics reveals the effect of maternal habitual dietary patterns on human amniotic fluid profile. *Scientific reports* **8**, 4076, doi:10.1038/s41598-018-22230-y (2018).
56 Holmes, E., Cloarec, O. & Nicholson, J. K. Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: application to HgCl2 toxicity. *Journal of proteome research* **5**, 1313-1320, doi:10.1021/pr050399w (2006).
57 Alves, A. C., Rantalainen, M., Holmes, E., Nicholson, J. K. & Ebbels, T. M. Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets. *Analytical chemistry* **81**, 2075-2084, doi:10.1021/ac801982h (2009).
58 Rodriguez-Martinez, A., Ayala, R., Posma, J. M. & Dumas, M. E. Exploring the Genetic Landscape of Metabolic Phenotypes with MetaboSignal. *Current protocols in bioinformatics* **61**, 14 14 11-14 14 13, doi:10.1002/cpbi.41 (2018).
59 Wang, Y. *et al.* Magic angle spinning NMR and 1H-31P heteronuclear statistical total correlation spectroscopy of intact human gut biopsies. *Analytical chemistry* **80**, 1058-1066, doi:10.1021/ac701988a (2008).
60 Keun, H. C. *et al.* Heteronuclear F-19-H-1 statistical total correlation spectroscopy as a tool in drug metabolism: Study of flucloxacillin biotransformation. *Analytical chemistry* **80**, 1073-1079, doi:10.1021/ac702040d (2008).
61 Tredwell, G. D., Bundy, J. G., De Iorio, M. & Ebbels, T. M. Modelling the acid/base (1)H NMR chemical shift limits of metabolites in human urine. *Metabolomics : Official journal of the Metabolomic Society* **12**, 152, doi:10.1007/s11306-016-1101-y (2016).
62 Aue, W. P., Karhan, J. & Ernst, R. R. Homonuclear Broad-Band Decoupling and 2-Dimensional J-Resolved Nmr-Spectroscopy. *J Chem Phys* **64**, 4226-4227, doi:Doi 10.1063/1.431994 (1976).
63 Nagayama, K., Kumar, A., Wuthrich, K. & Ernst, R. R. Experimental-Techniques of Two-Dimensional Correlated Spectroscopy. *J Magn Reson* **40**, 321-334, doi:Doi 10.1016/0022-2364(80)90255-3 (1980).
64 Aue, W. P., Bartholdi, E. & Ernst, R. R. 2-Dimensional Spectroscopy - Application to Nuclear Magnetic-Resonance. *J Chem Phys* **64**, 2229-2246, doi:Doi 10.1063/1.432450 (1976).
65 Bodenhausen, G. & Ruben, D. J. Natural Abundance N-15 Nmr by Enhanced Heteronuclear Spectroscopy. *Chem Phys Lett* **69**, 185-189, doi:Doi 10.1016/0009-2614(80)80041-8 (1980).
66 Keeler, J. *Understanding NMR Spectroscopy, Edn. 2 (John Wiley & Sons, Oxford, UK, 2002)*. 526
67 Bax, A., Farley, K. A. & Walker, G. S. Increased HMBC sensitivity for correlating poorly resolved proton multiplets to carbon-13 using selective or semi-selective pulses. *J Magn Reson Ser A* **119**, 134-138, doi:DOI 10.1006/jmra.1996.0063 (1996).
68 Bollard, M. E. *et al.* High-resolution (1)H and (1)H-(13)C magic angle spinning NMR spectroscopy of rat liver. *Magnetic resonance in medicine* **44**, 201-207 (2000).
69 Smith, L. M. *et al.* Statistical correlation and projection methods for improved information recovery from diffusion-edited NMR spectra of biological samples. *Analytical chemistry* **79**, 5682-5689, doi:10.1021/ac0703754 (2007).
70 Tang, H. R., Wang, Y. L., Nicholson, J. K. & Lindon, J. C. Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma. *Anal Biochem* **325**, 260-272, doi:10.1016/j.ab.2003.10.033 (2004).
71 Lenz, E. M. Nuclear magnetic resonance (NMR)-based drug metabolite profiling. *Methods in molecular biology* **708**, 299-319, doi:10.1007/978-1-61737-985-7_18 (2011).
72 Ramautar, R., Somsen, G. W. & de Jong, G. J. CE-MS in metabolomics. *Electrophoresis* **30**, 276-291, doi:10.1002/elps.200800512 (2009).

73 Garcia-Perez, I. *et al.* Metabolic fingerprinting of Schistosoma mansoni infection in mice urine with capillary electrophoresis. *Electrophoresis* **29**, 3201-3206, doi:10.1002/elps.200800031 (2008).
74 Fiehn, O. Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Current protocols in molecular biology* **114**, 30 34 31-30 34 32, doi:10.1002/0471142727.mb3004s114 (2016).
75 Spraul, M., Nicholson, J. K., Lynch, M. J. & Lindon, J. C. Application of the one-dimensional TOCSY pulse sequence in 750 MHz 1H-NMR spectroscopy for assignment of endogenous metabolite resonances in biofluids. *Journal of pharmaceutical and biomedical analysis* **12**, 613-618, doi:10.1016/0731-7085(93)e0028-l (1994).
76 Crockford, D. J. *et al.* Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. *Anal Chem* **78**, 363-371, doi:10.1021/ac051444m (2006).
77 Teul, J. *et al.* Improving metabolite knowledge in stable atherosclerosis patients by association and correlation of GC-MS and 1H NMR fingerprints. *Journal of proteome research* **8**, 5580-5589, doi:10.1021/pr900668v (2009).
78 Posma, J. M., Robinette, S. L., Holmes, E. & Nicholson, J. K. MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG. *Bioinformatics* **30**, 893-895, doi:10.1093/bioinformatics/btt612 (2014).
79 Quinn, R. A. *et al.* Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends in pharmacological sciences* **38**, 143-154, doi:10.1016/j.tips.2016.10.011 (2017).
80 Gratton, J. *et al.* Optimized Sample Handling Strategy for Metabolic Profiling of Human Feces. *Analytical chemistry* **88**, 4661-4668, doi:10.1021/acs.analchem.5b04159 (2016).
81 Farrant, R. D., Lindon, J. C. & Nicholson, J. K. Internal temperature calibration for 1H NMR spectroscopy studies of blood plasma and other biofluids. *NMR in biomedicine* **7**, 243-247 (1994).
82 Holmes, E. *et al.* 750 MHz 1H NMR spectroscopy characterisation of the complex metabolic pattern of urine from patients with inborn errors of metabolism: 2-hydroxyglutaric aciduria and maple syrup urine disease. *Journal of pharmaceutical and biomedical analysis* **15**, 1647-1659 (1997).
83 Duarte, I. F. *et al.* Identification of metabolites in human hepatic bile using 800 MHz 1H NMR spectroscopy, HPLC-NMR/MS and UPLC-MS. *Molecular bioSystems* **5**, 180-190, doi:10.1039/b814426e (2009).
84 Maaheimo, H., Rabina, J. & Renkonen, O. 1H and 13C NMR analysis of the pentasaccharide Gal beta (1-->4)GlcNAc beta (1-->3)-[GlcNAc beta (1-->6)]Gal beta (1-->4)GlcNAc synthesized by the mid-chain beta-(1-->6)-D-N-acetylglucosaminyltransferase of rat serum. *Carbohydrate research* **297**, 145-151 (1997).
85 Want, E. J. *et al.* Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* **5**, 1005-1018, doi:10.1038/nprot.2010.50 (2010).
86 Folch, J., Lees, M. & Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *The Journal of biological chemistry* **226**, 497-509 (1957).

CODE AND SOFTWARE AVAILABILITY

CA-PLS (and PLS, OSC-PLS): The code for executing the PLS, covariate-adjusted (O)PLS and simple orthogonal PLS/PLS-DA is provided in https://bitbucket.org/jmp111/capls/src/. This can be executed in a Matlab environment.

STORM (and STOCSY): The code for executing both the STOCSY and STORM algorithms is in https://bitbucket.org/jmp111/storm/src. These can be executed in a Matlab environment.

RED-STORM - The code for executing the RED-STORM algorithm is provided in https://bitbucket.org/jmp111/redstorm/src/. This can be executed in a Matlab environment.

RELATED LINKS

Key reference(s) using this protocol

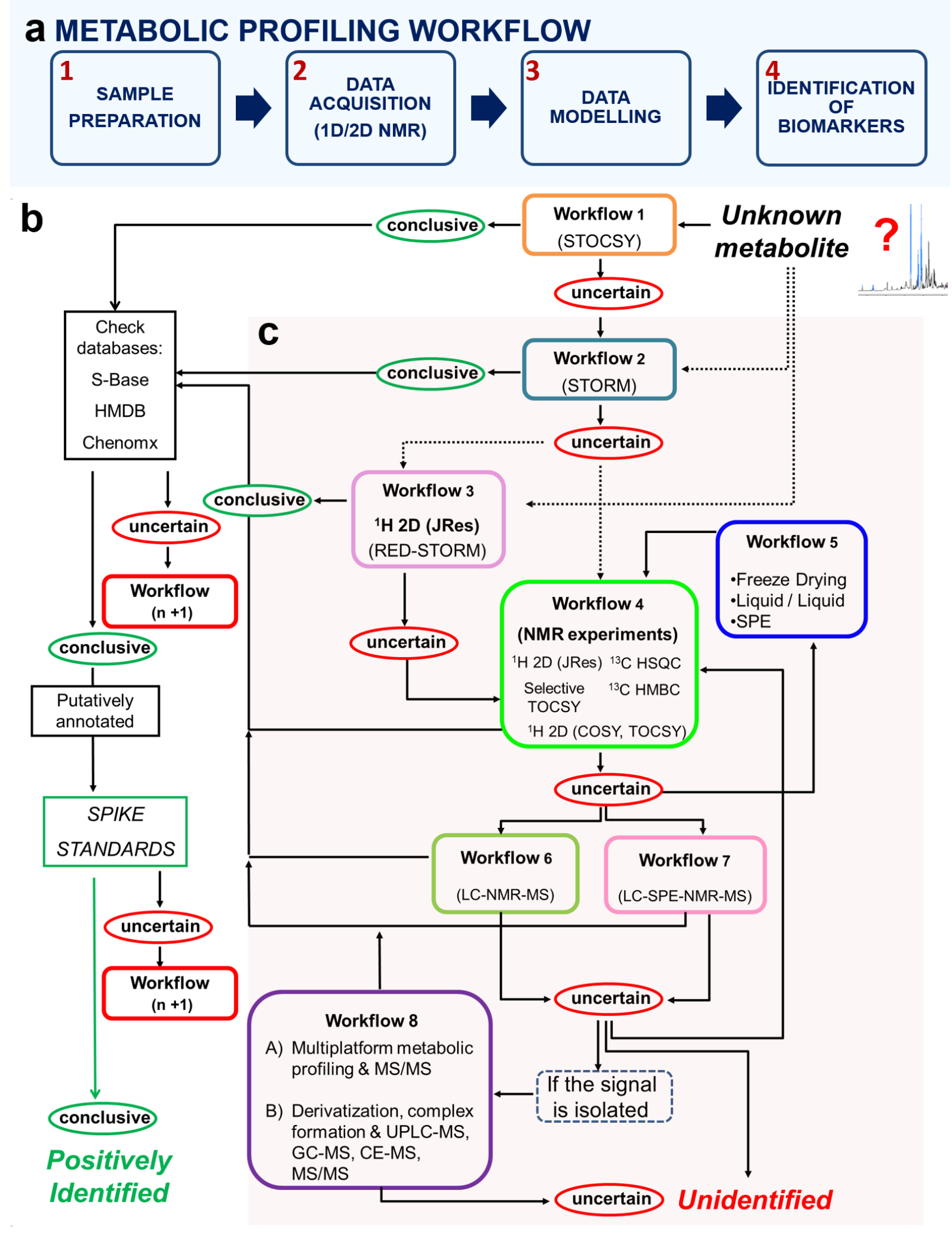
Elliott, P. *et al.* Urinary metabolic signatures of human adiposity. *Science translational medicine* **7**, 285ra262, doi:10.1126/scitranslmed.aaa5680 (2015).

Posma, J. M. *et al.* Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. *Anal Chem* **89**, 3300-3309, doi:10.1021/acs.analchem.6b03324 (2017).

Garcia-Perez, I. *et al.* Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. *Lancet Diabetes Endo* **5**, 184-195, doi:10.1016/S2213-8587(16)30419-3 (2017).

Key data used in this protocol (Optional heading)

Orbán-Németh, Z. et al. Nat. Protoc. 13, 478–494 (2018) [URL]


a METABOLIC PROFILING WORKFLOW
1 SAMPLE PREPARATION
2 DATA ACQUISITION (1D/2D NMR)
3 DATA MODELLING
4 IDENTIFICATION OF BIOMARKERS
b
Workflow 1 (STOCSY)
conclusive
Unknown metabolite
?
uncertain
Check databases:
S-Base
HMDB
Chenomx
c
Workflow 2 (STORM)
conclusive
uncertain
Workflow 3
1H 2D (JRes)
(RED-STORM)
conclusive
uncertain
Workflow (n +1)
Workflow 5
•Freeze Drying
•Liquid / Liquid
•SPE
Workflow 4
(NMR experiments)
1H 2D (JRes) 13C HSQC
Selective TOCSY 13C HMBC
1H 2D (COSY, TOCSY)
uncertain
conclusive
Putatively annotated
SPIKE STANDARDS
uncertain
Workflow 6
(LC-NMR-MS)
Workflow 7
(LC-SPE-NMR-MS)
uncertain
Workflow (n +1)
uncertain
Workflow 8
A) Multiplatform metabolic profiling & MS/MS
B) Derivatization, complex formation & UPLC-MS, GC-MS, CE-MS, MS/MS
If the signal is isolated
conclusive
Positively Identified
uncertain
Unidentified

1H-NMR data acquisition

↓

STOCSY

↓

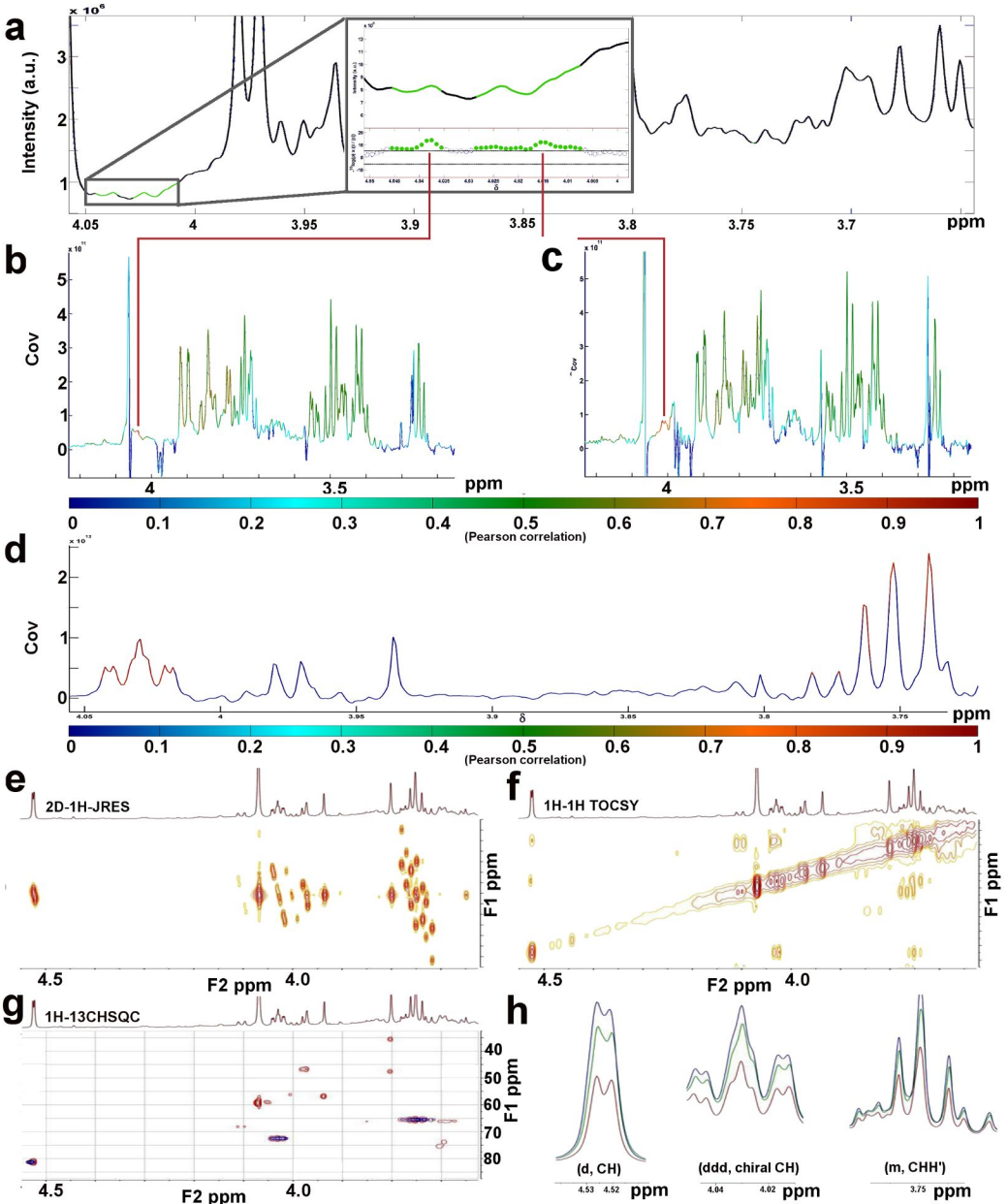
STORM

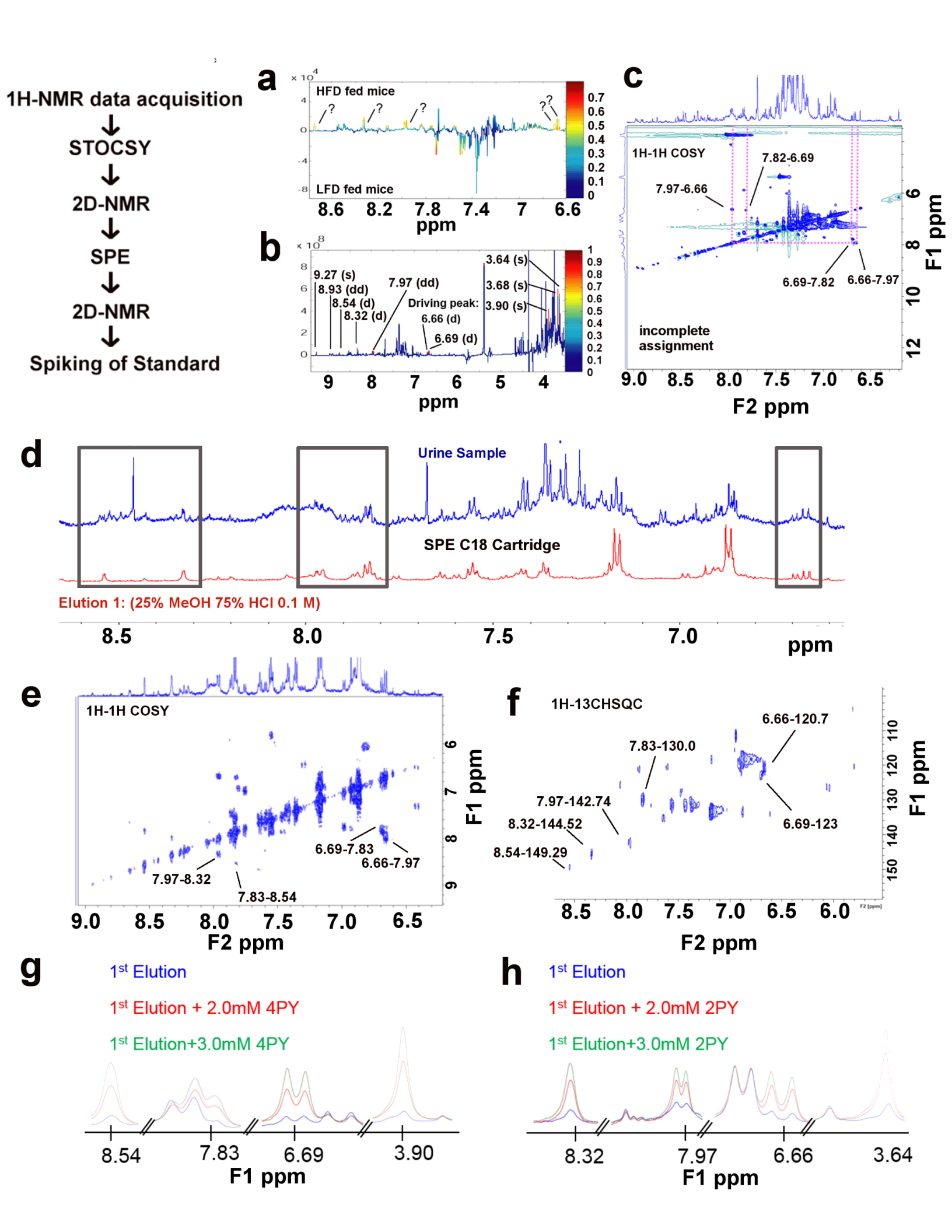
↓

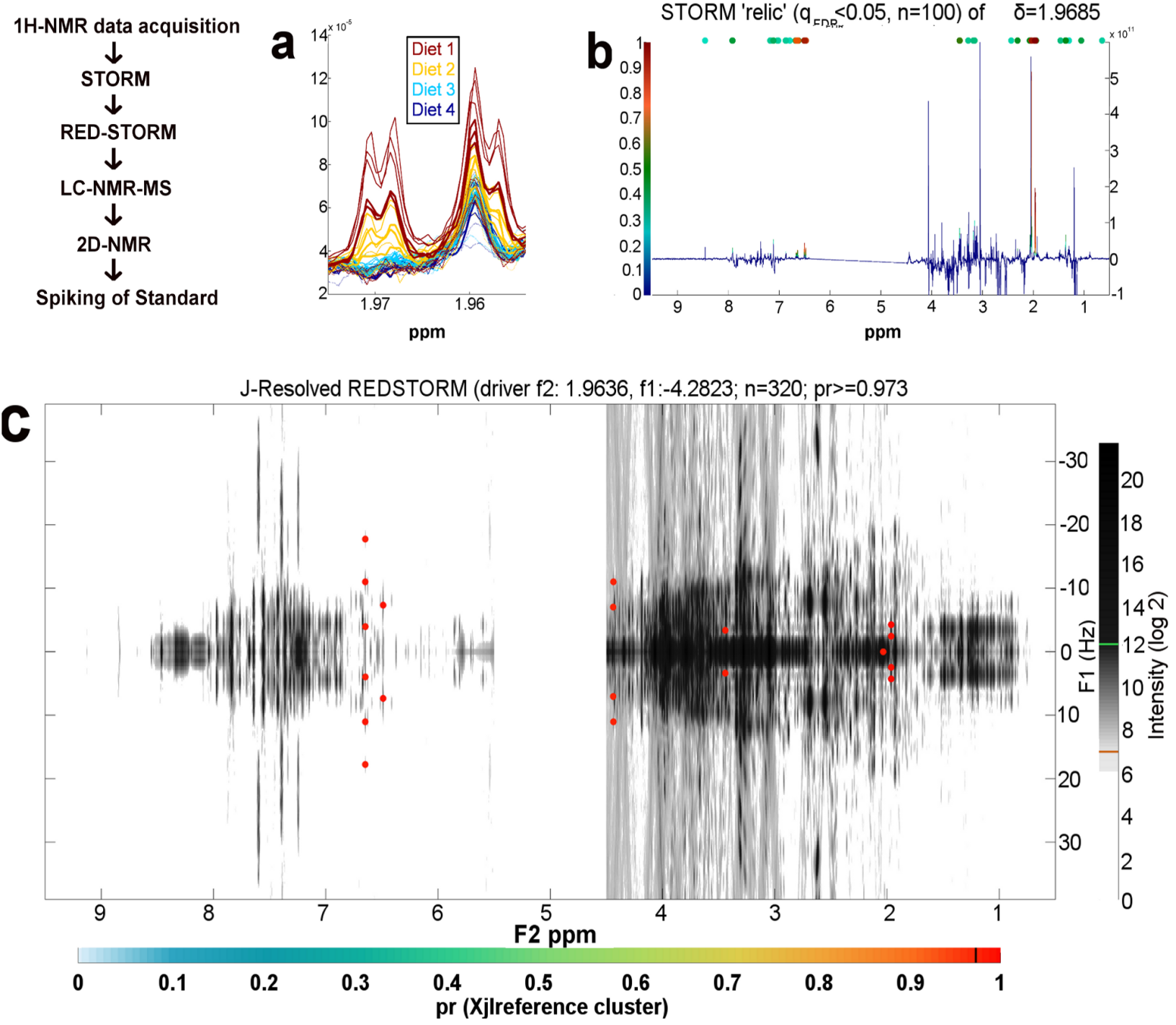
2D-NMR

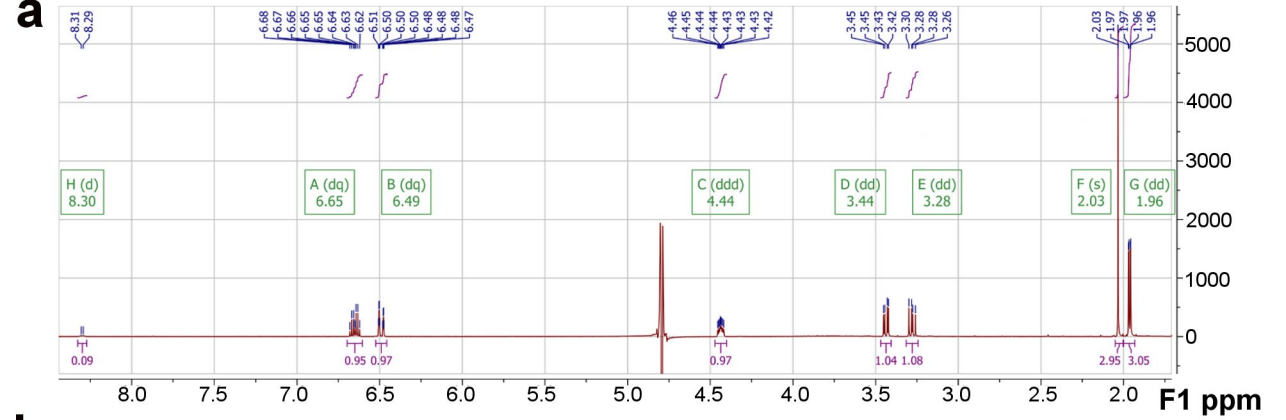
↓

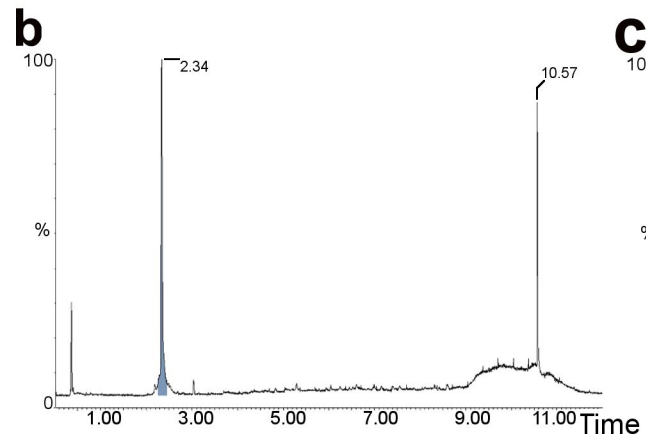
Spiking of Standard

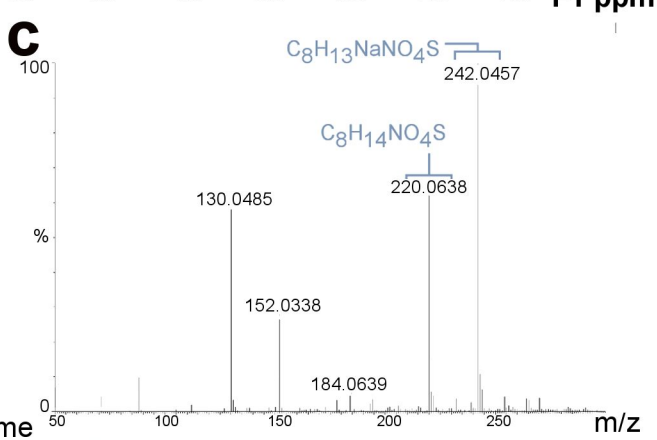



1H-NMR data acquisition
STOCSY
2D-NMR
SPE
2D-NMR
Spiking of Standard
a
HFD fed mice
LFD fed mice
ppm
b
9.27 (s)
8.93 (dd)
8.54 (d)
8.32 (d)
7.97 (dd)
Driving peak:
6.66 (d)
6.69 (d)
3.64 (s)
3.68 (s)
3.90 (s)
ppm
c
1H-1H COSY
7.82-6.69
7.97-6.66
6.69-7.82
6.66-7.97
incomplete assignment
F2 ppm
F1 ppm
d
Urine Sample
SPE C18 Cartridge
Elution 1: (25% MeOH 75% HCl 0.1 M)
ppm
e
1H-1H COSY
7.97-8.32
7.83-8.54
6.69-7.83
6.66-7.97
F2 ppm
F1 ppm
f
1H-13CHSQC
6.66-120.7
7.83-130.0
7.97-142.74
8.32-144.52
8.54-149.29
6.69-123
F2 ppm
F1 ppm
g
1st Elution
1st Elution + 2.0mM 4PY
1st Elution+3.0mM 4PY
F1 ppm
h
1st Elution
1st Elution + 2.0mM 2PY
1st Elution+3.0mM 2PY
F1 ppm


1H-NMR data acquisition
STORM
RED-STORM
LC-NMR-MS
2D-NMR
Spiking of Standard
a
Diet 1
Diet 2
Diet 3
Diet 4
ppm
b
STORM 'relic' (q_FDR<0.05, n=100) of δ=1.9685
ppm
c
J-Resolved REDSTORM (driver f2: 1.9636, f1:-4.2823; n=320; pr>=0.973
F2 ppm
F1 (Hz)
Intensity (log 2)
pr (Xjlreference cluster)



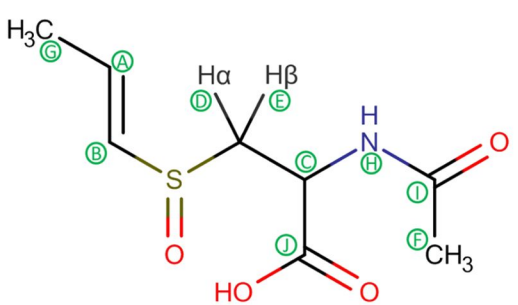




d

Peak	δ (^{1}H)	Multiplicity	Integral	J-coupling	TOCSY	COSY	δ (^{13}C)	HSQC	HMBC	Type
A	*6.65*	dq	0.95	*15.22, 6.78*	A, B, G	A, B, G	*146.46*	A, B, G	B, G	**-CH=**C(-)-
B	*6.49*	dq	0.97	*15.21, 1.50*	A, B, G	A, B, G	*132.93*	A, B	A, G	**-CH=**C(-)-
C	*4.44*	ddd	0.97	*9.78, 8.20, 4.21*	C, D, E, H	C, D, E, H	*52.70*	C	J	-$CH_\alpha H_\beta$-**CH**(-NH)-
D	*3.44*	dd	1.04	*13.38, 4.21*	C, D, E, H	C, D, E	*57.67*	D	C, J	-**$CH_\alpha H_\beta$**-
E	*3.28*	dd	1.08	*13.38, 9.91*	C, D, E, H	C, D, E	*57.69*	E	C, J	-**$CH_\beta H_\alpha$**-
F	*2.03*	s	2.95	-	F	F	*24.85*	F	I	-**CH_3**
G	*1.96*	dd	3.05	*6.78, 1.60*	A, B, G	A, B, G	*20.48*	A, G	A, B, D, E	-**CH_3**
H	*8.30*	d	0.09	*8.30*	C, D, E, H	C, H	-	-	-	**-NH-**
I	-	-	-	-	-	-	*176.22*	-	-	-NH-**C(=O)**-C-
J	-	-	-	-	-	-	*178.05-178.24*	-	-	HO-**C(=O)**-C-

e



N-acetyl-S-(1Z)-propenyl-cysteine-sulfoxide (NAcSPCSO)