Accurate, fully-automated NMR spectral profiling for metabolomics

Siamak Ravanbakhsh[1,2], Philip Liu[1,3], Trent C. Bjorndahl[1,3], Rupasri Mandal[1,3], Jason R. Grant[1], Michael Wilson[1], Roman Eisner[1], Igor Sinelnikov[3], Xiaoyu Hu[4], Claudio Luchinat[5], Russell Greiner[1,2], and David S. Wishart[1,3,6]

[1]*Department of Computing Science, University of Alberta, Edmonton, AB, Canada*
[2]*Alberta Innovates Center for Machine Learning (AICML)*
[3]*Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada*
[4]*Fiorgen Foundation, 50019 Sesto Fiorentino, Italy*
[5]*Centro Risonanze Magnetiche (CERM/CIRMMP), University of Florence, 50019 Sesto Fiorentino, Italy*
[6]*National Research Council, National Institute for Nanotechnology (NINT)*

Abstract—Many diseases cause significant changes to the concentrations of small molecules (*a.k.a.* metabolites) that appear in a person's biofluids, which means such diseases can often be readily detected from a persons metabolic profile.This information can be extracted from a person's biofluids using NMR spectroscopy. Today, this is often done manually by trained experts, which means this process is relatively slow, expensive and can be error-prone. A system that can quickly, accurately and autonomously produce a person's metabolic profile would enable efficient and reliable prediction of many such diseases from a single sample, which could significantly improve the way medicine is practiced.

This paper presents such a system: Given a 1D ^{1}H NMR spectrum of a complex biofluid such as serum or CSF, our BAYESIL system can automatically determine this metabolic profile, and do so without any human guidance. This requires first performing all of the required spectral processing steps (*i.e.*, Fourier transformation, phasing, solvent-removal, chemical shift referencing, baseline correction, lineshape convolution) then matching this resulting spectrum against a reference compound library, which contains the signatures of each relevant metabolite. Many of these processing steps are novel algorithms, and our matching step views spectral matching as an inference problem within a probabilistic graphical model that rapidly approximates the most probable metabolic profile.

Our extensive studies on a diverse set of complex mixtures (real biological samples, defined mixtures and realistic computer generated spectra; each involving ~ 50 compounds), show that BAYESIL can autonomously and accurately find NMR-detectable metabolites at concentrations as low as 2μM, in terms of both identification ($\sim 90\%$ correct) and quantification ($\sim 10\%$ error), in under 5 minutes on a single CPU processor. These results demonstrate that BAYESIL is the first fully-automatic publicly-accessible system that provides quantitative NMR spectral profiling effectively – with an accuracy that meets or exceeds the performance of highly trained human experts. We anticipate this tool will usher in high-throughput metabolomics and enable a wealth of new applications of NMR in clinical settings. Users can access BAYESIL at *http://www.bayesil.ca*.

Index Terms— NMR — Metabolomics — Targeted Profiling — Probabilistic Graphical Models NMR — Metabolomics — Targeted Profiling — Probabilistic Graphical Models—

Metabolomics is a relatively new branch of "omics" science that focuses on the system-wide characterization of metabolites [1]. Metabolomics is often viewed as complementary to the other "omics" fields as it provides information about both an organism's phenotype and its environment. Because metabolomics provides a unique window on gene-environment interactions, it is playing an increasingly important role in many quantitative phenotyping and functional genomics studies [2], [3], [4]. It is also finding more applications in disease diagnosis, biomarker discovery and drug development/discovery (*e.g.*, [5], [6], [7]). This rapid growth in interest and excitement surrounding metabolomics is also revealing its "Achilles heel": Unlike proteomics, genomics or transcriptomics, which are *high-throughput* sciences, metabolomics is a relatively *low-throughput* science. Compared to genomics, where it is now possible to automatically characterize 1000s of genes, 100s of thousands of transcripts and millions of SNPs in mere minutes, metabolomics only allows users to identify and measure a few dozen metabolites after many hours of manual effort. In other words, *metabolomics is not automated*. This problem may stem from the history of metabolomics, as its analytical techniques, such as NMR spectroscopy, gas-chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS), were originally developed for identifying and quantifying *pure* compounds, not complex mixtures. Because most biological samples contain hundreds of metabolites, the resulting NMR, HPLC or LC-MS spectra usually contain hundreds or even thousands of peaks. The challenge in metabolomics, therefore, is to identify the mixture of compounds that produced this forest of peaks. This compound identification process, called *spectral profiling*, involves fitting the mixture spectrum to a set of individual pure reference spectra obtained from known compounds [8], [9]. If done correctly, the fitting process yields not only the identity of the compounds, but also the concentration of those compounds. Therefore, the end result of a successful spectral profiling study is a table of metabolite names and their absolute or relative concentrations. Because spectral profiling is such a complex pattern recognition problem, it is often best done by a trained expert. However, this reliance on manual data analysis by a human expert is problematic, as it is slow and leads to inconsistent results, operator errors and reduced levels of reproducibility [10].

The automation bottleneck in metabolomics is widely recognized, and has led to a number of efforts to accelerate or automate compound identification and/or quantification in LC-MS, in GC-MS and in NMR spectroscopy. Some of the most active efforts in (semi)automated compound identification and quantification have been in NMR-based metabolomics. In particular, several software packages have been developed that support semi-automatic NMR spectral profiling of 1D and 2D ^{1}H NMR spectra, including some commercial packages (*e.g.*, [12], [14], [13]). There are tremendous differences in the quantification/identification capabilities, spectral database size, speed and instrument compatibility of these software tools. Furthermore, these packages either require manual fitting or manual spectral processing, or a bit of both;
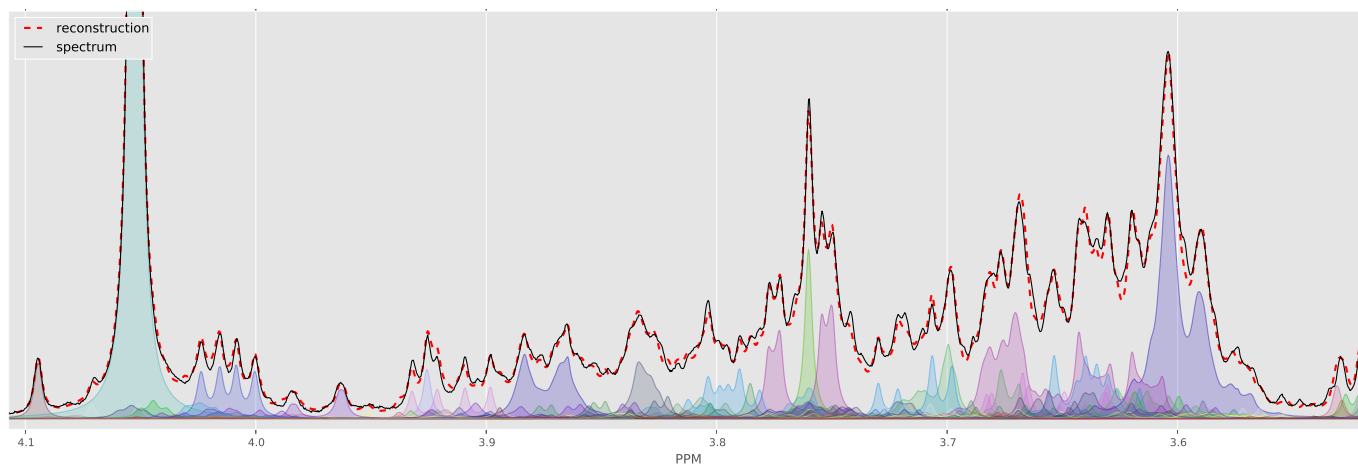
Fig. 1. This figure shows the crowded region (3.5-4.1 PPM) of a computer generated spectrum with 150 compounds (solid black) and the fit produced by BAYESIL (dashed red) as well as individual clusters as quantified by BAYESIL. Each cluster is free to shift a specified amount, which is at least 0.025 PPM.

see Appendix I for a comprehensive list of NMR softwares and their limitations. The need for such manual interventions leads to a number of issues, including slower throughput, operator fatigue and associated operator errors, the need for highly trained and dedicated experts, the requirement of two or more spectral assessments for quality assessment and control purposes, and inconsistent results between individuals, between labs or over different time periods [8], [10].

It would be better to have a software system that can automatically perform both spectral processing and spectral deconvolution, be able to analyze complex mixtures ($\sim$60 compounds) quickly and accurately, and be able to produce reliable compound concentrations. Here we describe such a system, called BAYESIL.

Extensive testing, on computer-generated and laboratory-generated chemical mixtures as well as real biological samples, shows that BAYESIL consistently performs with $\sim 90\%$ accuracy for compound identification in mixtures with up to 60 different metabolites. It also determines metabolite concentrations with $\sim 10\%$ error. For computer-generated biofluid spectra, where the ground truth is known, BAYESIL consistently outperforms highly trained experts in both identification and quantification. BAYESIL appears to be the first system that supports fully automated and fully quantitative NMR-based metabolomics. This paper describes this system, its underlying algorithms and its performance across various tests.

I. METABOLIC PROFILING PIPELINE

BAYESIL performs fully automated spectral processing and spectral profiling for 1D ^{1}H NMR spectra collected on either Agilent/Varian or Bruker instruments, at several different frequencies. In particular, it uses a variety of intelligent phasing and baseline correction methods to automatically process raw 1D NMR spectra (*i.e.*, FIDs). It also uses approximate inference techniques to rapidly perform very accurate spectral deconvolution, yielding both compound identities and their concentrations. Here we briefly describe BAYESIL's spectral processing algorithms, the principles and rationale behind BAYESIL's spectral deconvolution method and the construction of BAYESIL's spectral library.

A. Spectral Processing in BAYESIL

Successful NMR spectral profiling depends critically on the quality and uniformity of the starting NMR spectrum. Unfortunately, most spectral processing functions (*i.e.*, phasing, baseline correction, solvent filtering, chemical shift referencing) are left

to the user. Given the complexity and large number of variables, values and filters that can be used, many view spectral processing more as an art, rather than a science. Different perspectives or different personal thresholds on what is a "good looking" NMR spectrum can potentially lead to very different results regarding what compounds are identified or which compounds are accurately quantified in a biofluid spectrum. To address this issue, BAYESIL itself performs all of the spectral processing functions: starting from the FID, it performs zero-filling, Fourier and Hilbert transformation, phasing, baseline correction, chemical shift referencing, reference deconvolution and smoothing. Automating this process ensures reproducibility, consistency and uniformity of the input data prior to spectral deconvolution. Here we briefly sketch some of the more challenging steps in this process.

Phasing involves maximizing the symmetry of the peaks by reducing zero-order and first-order phase mismatch. Zero-order phase mismatch is a sign of the difference between the reference phase and the receiver phase and is independent of frequency. The first-order phase mismatch can be a result of the time-delay between excitation and detection, flip-angle variation and the filter that is used to reduce the noise outside of the spectral bandwidth [15]. In addition to using well-known techniques, such as spectral norm minimization [16], BAYESIL uses the cross entropy optimization method [11] to jointly maximize a direct measure of peak symmetry for isolated peaks across the spectrum.

Baseline correction involves removing distortions that may arise from hardware artifacts or highly concentrated components of the mixture (*e.g.*, solvent), while keeping the desirable signal intact. This process is often performed in two steps: 1) baseline-detection and 2) modelling. BAYESIL relies on iterative thresholding [17] and estimating the signal-to-noise ratio to detect the baseline points. It uses monotonic cubic Hermite interpolation [18] and Whittaker smoothing technique for baseline modelling [19].

BAYESIL also provides the options for *smoothing and line-broadening* using Savitzky-Golay [20] and Gaussian filters. However smoothing is mostly cosmetic and it is not essential for spectral deconvolution. In fact, it may degrade the signal and occasionally remove the the low-amplitude and narrow peaks. Similarly, we found the effect of *reference deconvolution* – which may be used to remove instrumental or experimentally induced distortions of the Lorentzian lineshape – is also mostly cosmetic, and if the distortion around the reference peak has any source other than poor shimming, using reference deconvolution will have an adverse effect on the rest of the NMR spectrum.

B. Spectral Deconvolution

An NMR spectrum for a compound $\mathcal{M}$ is a collection of one or more *Lorentzian* peaks formed into one or more clusters – that is, each compound $\mathcal{M}$ is a set of clusters $\{c_k\}$, where each cluster c_k is set of peaks, and each peak is defined by a triple, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ corresponding to its height, center and width (at half height) respectively – where the height at x due to this peak is $\mathsf{q}(x; \boldsymbol{\theta}) = \frac{\theta_1 \theta_3}{\theta_3 + 4(\theta_2 - x)^2}$. Letting $\mathcal{X}$ refer to the entire spectrum (*e.g.*, from -1 to 13 PPM when referenced against the DSS peak), the height of the spectrum of a pure compound $\mathcal{M}$ at each location $x \in \mathcal{X}$, is

$$\widehat{\mathsf{s}}(x; \mathcal{M}, \rho_{\mathcal{M}}, \boldsymbol{\delta}_{\mathcal{M}}) \quad = \quad \rho_{\mathcal{M}} \sum_{\mathcal{C} \in \mathcal{M}} \sum_{\boldsymbol{\theta} \in \mathcal{C}} \mathsf{q}(x - \delta_{\mathcal{C}}; \boldsymbol{\theta}) \qquad (1)$$

where $\rho_{\mathcal{M}}$ is the concentration of this compound and $\boldsymbol{\delta}_{\mathcal{M}} = \{\delta_{\mathcal{C}} \mid c \in \mathcal{M}\}$ is the set of chemical shifts for the clusters associated with this compound.

An NMR spectrum is essentially a linear combination of the peaks in its component compounds: that is, the height at each PPM value x of a mixture spectrum is just the sum of the contributions of each compound. This means, given the concentrations of the compounds $\boldsymbol{\rho} = \{\rho_{\mathcal{M}}\}$, and the chemical shifts $\boldsymbol{\delta} = \bigcup_{\mathcal{M}} \boldsymbol{\delta}_{\mathcal{M}}$ of the clusters associated with these compounds, we can then "draw" an NMR spectrum – *i.e.*, the height at each PPM x, given this $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$, is $\widehat{\mathsf{s}}(x; \boldsymbol{\rho}, \boldsymbol{\delta}) = \sum_{\mathcal{M}} \widehat{\mathsf{s}}(x; \mathcal{M}, \rho_{\mathcal{M}}, \boldsymbol{\delta}_{\mathcal{M}})$.

The spectral deconvolution challenge, in general, is the reverse process: Given a set of compounds $\{\mathcal{M}_1, \ldots, \mathcal{M}_r\}$ with associated signatures (*i.e.*, $\boldsymbol{\theta}$ values of their peaks, organized in clusters) and the observed spectrum $\mathsf{s}(\cdot)$, find the "best" combination of concentrations $\boldsymbol{\rho}$ and shifts $\boldsymbol{\delta}$ to fit that spectrum. To determine which values are best, for now, we consider a simple loss function that is the (square of the) difference of the heights between the observed spectrum $\mathsf{s}(\cdot)$ and the reconstructed spectrum $\widehat{\mathsf{s}}(\cdot; \boldsymbol{\delta}, \boldsymbol{\rho})$

$$\ell_{\mathcal{X}}(\,\mathsf{s}(\cdot),\, \widehat{\mathsf{s}}(\cdot; \boldsymbol{\rho}, \boldsymbol{\delta})\,) \;=\; \int_{x \in \mathcal{X}} \big(\mathsf{s}(x) - \widehat{\mathsf{s}}(x; \boldsymbol{\rho}, \boldsymbol{\delta})\big)^2 \, \mathrm{d}x \qquad (2)$$

where the subscript $\mathcal{X}$ indicates that this loss function applies to the entire spectrum; see Appendix II for BAYESIL's actual loss function.

Our task is to find the values of

$$[\boldsymbol{\rho}^*,\ \boldsymbol{\delta}^*] \quad = \quad \arg_{\boldsymbol{\rho}, \boldsymbol{\delta}} \min \quad \ell_{\mathcal{X}}(\mathsf{s}(\cdot), \widehat{\mathsf{s}}(\cdot; \boldsymbol{\rho}, \boldsymbol{\delta})) \qquad (3)$$

that minimize this loss function. Figure 1 shows part of a spectrum over a complex mixture, and BAYESIL's solution obtained by minimizing the loss function.

This corresponds to search over a huge space – all possible shifts for each of the clusters, and all possible concentrations over the compounds. The key innovation of BAYESIL is how it minimizes this highly non-linear loss function, efficiently. In particular, BAYESIL "factors" this large task into a set of inter-related smaller tasks. Two characteristics of the NMR spectra make this factorization possible: 1) each shift is over only a small range (typically a window of ± 0.025 PPM); and 2) as the height in the spectrum due to a Lorentzian peak diminishes quadratically from its center, each peak and therefore each cluster can only "influence" a small interval; see Appendix III.

Now consider a function that maps each point in the spectrum to the set of clusters that might affect the height at this location. That is, given the set of compounds $\{\mathcal{M}\}$ that might appear in a particular biofluid (*e.g.*, the 48 that can appear in CSF), we can identify each PPM location $x \in \mathcal{X}$ with the small set of clusters that might influence it, $c(x)$.

We can then partition the spectrum into disjoint contiguous regions, $\{\mathcal{X}_I\}$, where every PPM location in each $\mathcal{X}_I$ involves exactly the same subset of clusters – *i.e.*, for any pair of points $x_1, x_2 \in \mathcal{X}_I$, we know that $c(x_1) = c(x_2)$. For example, our library for CSF includes 48 compounds, with a total of 180 clusters and 946 peaks. A typical CSF spectrum is partitioned into
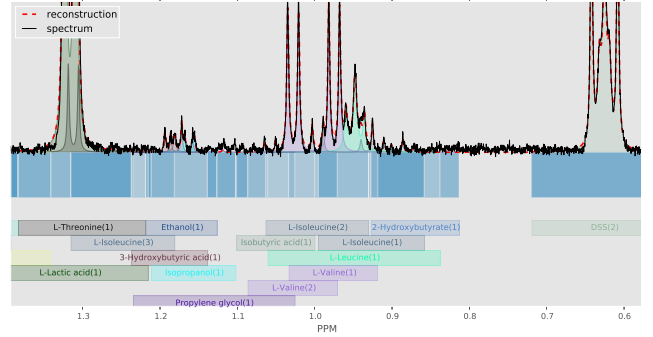

Fig. 2. Partitioning of $\mathcal{X}$ into continuous blocks $\mathcal{X}_I \subset \mathcal{X}$ for a part of a human serum spectrum. Here each block is shown with a different shade of blue, below the horizontal axis. The domain of influence of each cluster is also indicated with coloured blocks.

$\sim$350 regions. Each of these regions will involve between 1 and $\sim$25 clusters, and span between 0.0001 and 1.5 PPM. Moreover, each cluster will appear in between 1 and $\sim$70 different regions.

Figure 2 shows the division of a part of human serum NMR spectrum into regions $\mathcal{X}_I$; blocks in different shades of blue. For example the region $\mathcal{X}_{[.8563, .9289]}$ from 0.8563 to 0.9289 PPM might include significant contributions from the first cluster of 2-Hydroxybutyrate, the first cluster of L-Isoleucine and/or the first cluster of L-Leucine. The region immediately to its left (from 0.9289 to 0.9370 PPM) includes these and also a cluster of L-Valine, and the one to the right (from 0.8563 to 0.87526 PPM) does not include L-Isoleucine.

As the loss function $\ell(\cdot, \cdot)$ is additive over the domain $\mathcal{X}$, we can rewrite the optimization of eq[3] as the sum of the losses for each of the regions $\mathcal{X}_I$:

$$[\boldsymbol{\rho}^*, \boldsymbol{\delta}^*] \;=\; \arg_{\boldsymbol{\rho}, \boldsymbol{\delta}} \min \sum_I \ell_{\mathcal{X}_I}(\mathsf{s}(\cdot), \widehat{\mathsf{s}}(\cdot; \boldsymbol{\rho}_I, \boldsymbol{\delta}_I)) \qquad (4)$$

Now recall that each region $\mathcal{X}_I$ involves relatively few compounds and clusters. This suggests a preliminary step of simply "solving" each region, by itself: *i.e.*, find the best centers for the clusters in that region $\boldsymbol{\delta}_I$, and the best concentrations for the associated compounds $\boldsymbol{\rho}_I$, which collectively minimize the loss over the PPM-interval $\mathcal{X}_I$. This simple approach is fast, as it involves relatively few variables and a limited range of PPM-values. Unfortunately, this does not produce the overall correct answer – that is, each region has an opinion about the concentration and shift values of its cluster, and when two (or more) regions each involve the same variable, they must both agree on its value.

To address this problem, we take a probabilistic approach, viewing the task of minimizing the loss function eq[2] as finding the "Maximum a Posteriori" (MAP) assignment – *i.e.*, the assignment to all of the cluster-shift and compound-concentration $[\boldsymbol{\delta}, \boldsymbol{\rho}]$ variables that makes the observed data as likely as possible. Here, the Boltzmann formula gives the probabilistic interpretation of the loss (*a.k.a.* the energy)

$$\mathbb{P}(\boldsymbol{\rho}, \boldsymbol{\delta} \mid \mathsf{s}(\cdot)) = \frac{1}{Z} \exp\left\{ -\frac{1}{T} \ell_{\mathcal{X}}(\mathsf{s}(\cdot), \widehat{\mathsf{s}}(\cdot; \boldsymbol{\rho}, \boldsymbol{\delta})) \right\} \qquad (5)$$

where Z is the normalization constant and T is known as the "temperature" parameter, explained in Appendix IV. Using the decomposition of loss over regions (eq[4]) we can write this distribution in factored form

$$\mathbb{P}(\boldsymbol{\rho}, \boldsymbol{\delta} \mid \mathsf{s}(\cdot)) = \frac{1}{Z} \prod_I \mathsf{f}_I(\boldsymbol{\rho}_I, \boldsymbol{\delta}_I)$$

$$\mathsf{f}_I(\boldsymbol{\rho}_I, \boldsymbol{\delta}_I) = \exp\left\{ -\frac{1}{T} \ell_{\mathcal{X}_I}(\mathsf{s}(\cdot), \widehat{\mathsf{s}}(\cdot; \boldsymbol{\rho}_I, \boldsymbol{\delta}_I)) \right\} \qquad (6)$$
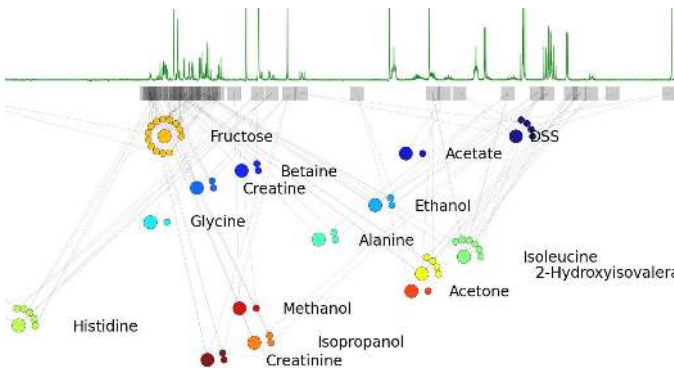
Fig. 3. Factor-graph for a library of 15 compounds immediately below an associated NMR spectrum. Each factor is represented by a square and each variable using a circle. Concentration (larger circles) and shift variables (smaller circles, beside the associated concentration) corresponding to each compound appear together. The position of each factor f_I position in the plot corresponds to the center of the corresponding block $\mathcal{X}_I$.
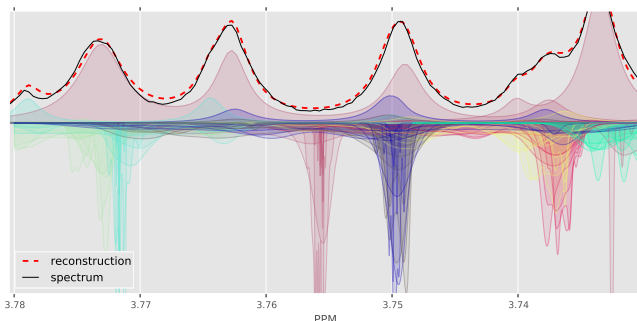

Fig. 4. A small region of human serum spectrum. The plots above horizontal axis show the original spectrum (solid black), individual clusters as well as overall fit (dashed red). The curves below horizontal axis show the BAYESIL's distribution over chemical shift variables for each cluster (c), over 6 iterations of spectral deconvolution. The distributions become more peaked towards the correct center in each iteration. Distributions below the horizon have the color of their associated cluster.

This decomposition of the distribution $\mathbb{P}(\rho, \delta \,|\, s(\cdot))$ can be represented using a probabilistic graphical model, known as a factor graph [21], which is a graph with two types of nodes: 1) factors (corresponding to regions or f_I), and 2) variables (here, concentrations and chemical shifts). Each factor has arcs that point only to its associated variables. Figure 3 shows a portion of this factor-graph for a simple defined mixture of 15 compounds.

By formulating the spectral deconvolution problem as MAP inference in a factor-graph, we have a variety of inference techniques at our disposal [21]. BAYESIL uses a non-parametric sequential Monte Carlo method [22] tailored to this inference problem (see Appendix IV for details). Using a Gaussian distribution around a set of particles (*a.k.a.* kernel density estimation) BAYESIL represents a distribution over each concentration $\rho_{\mathcal{M}}$ and shift variable $\delta_{\mathcal{C}}$. By reducing the temperature parameter, these distributions are gradually narrowed in each iteration until convergence, at which point the mode of the distributions represent an approximate MAP assignment. Figure 4 shows the evolution of distributions over the chemical shift variables over 6 iterations of spectral deconvolution.

C. BAYESIL's Spectral Library

Key to the success of any spectral deconvolution algorithm is the quality and size of its spectral library. We therefore collected 1D ^{1}H NMR reference spectra for each of the compounds in

BAYESIL's spectral library using pure compounds obtained from the Human Metabolome Library [1], using a standard protocol (see Materials and Methods). The spectral library contains relevant information about each compound ($\mathcal{M}$) including individual peak clusters (c) and peak amplitude positions and widths (*i.e.*, θ in eq[1]), as well as allowable chemical shift window $\underline{\delta}_{\mathcal{C}} \leq \delta_{\mathcal{C}} \leq \overline{\delta}_{\mathcal{C}}$ for each cluster c.

To analyze each biofluid, BAYESIL uses a specific spectral sub-library – here, one for serum and another one for CSF. The serum library consists of 50 NMR-detectable compounds from the human serum metabolome [23] while the CSF library consists of the 48 NMR-detectable compounds from the human CSF metabolome [24]; see Appendix V. The use of biofluid-specific or organism-specific spectral libraries significantly improves the performance of the spectral fitting process as it reduces the number of possible explanations for each peak.

II. ASSESSMENT

BAYESIL was assessed using 3 different types of spectral data sets over two different types of biofluids:

(a) Computer generated mixtures derived from its spectral library. We generated 5 random serum and 5 random CSF spectra by sampling from the distribution of the measured concentration ranges of various compounds, and the probability of observing them in the mixture from [23], [24]. The chemical shifts were also randomly sampled according to the chemical shift ranges from the corresponding spectral libraries. To be more realistic, we also added a small amount of noise to the spectrum – independently to the height at each position $x \in \mathcal{X}$. These correspond to "perfect" spectra, and are intended to assess the robustness and performance limits of BAYESIL under optimal conditions.

(b) Defined mixtures prepared in the laboratory. We created 15 defined mixtures (5 defined mixture of serum, 5 defined mixture of CSF, 5 random mixture of compounds in both serum and CSF, involving > 60 compounds), using carefully measured pure compounds and freshly prepared solutions. These provide real spectral data that probably include common spectral and solution artifacts (baseline and phasing issues, minor spontaneous reaction products, contaminants, matrix or pH effects). This set was used to assess BAYESIL's performance under well-controlled conditions where the composition and of the mixtures was almost perfectly known.

(c) Biological serum and CSF samples. We took human CSF and serum samples from previously studied samples that had been analyzed and quantified by NMR experts – here, 50 human serum and 5 human CSF samples. The set of compound mixtures was used to assess BAYESIL's performance under realistic conditions with common spectral and solution artifacts. Although human CSF contains a smaller number of NMR-detectable compounds than human serum, it is more difficult to profile due to the lower concentration of metabolites. While both the biological samples and defined mixtures were thoroughly analyzed, their exact compound concentrations cannot be perfectly known.

Overall, we believe these 3 test sets provide a robust assessment of BAYESIL's performance (as well as its limitations) under a wide range of conditions.

BAYESIL estimates the "detection threshold" based on the signal to noise ratio (SNR) in each spectrum – *i.e.*, when the signal is noisy this threshold is increase to provide a more confident identification and quantification. The SNR and therefore the detection threshold is directly related to the number of scans during spectral acquisition. For example our biological serum samples are produced using 128 scans and therefore most detection thresholds are $\sim 10\mu$M. Since compound concentrations in our CSF samples are considerably lower than serum, their analysis requires higher quality spectra (see Materials and Methods). Our CSF samples use 1024 scans resulting in detection thresholds that are often less than 2μM. However this threshold

TABLE I
IDENTIFICATION AND QUANTIFICATION ACCURACY OF BAYESIL AND HUMAN EXPERT ON VARIOUS DATA-SETS.

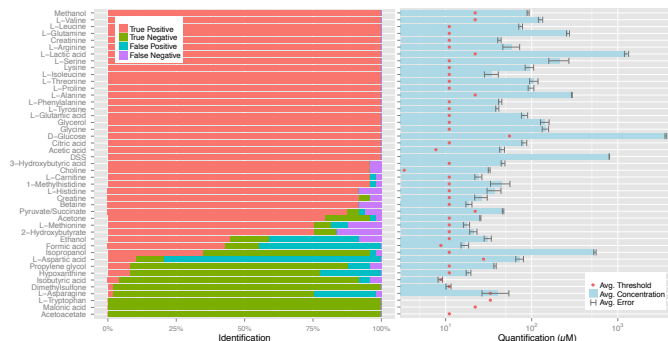| | | serum | | | CSF | | | complex |
		biological	def. mix.	comp. gen.	biological	def. mix.	comp. gen.	def. mix.
BAYESIL	id. accuracy	$.93 \pm .04$	$.94 \pm .02$	$.98 \pm .01$	$.90 \pm .04$	$.89 \pm .03$	$.95 \pm .03$	$.90 \pm .02$
	quant. accuracy	$.89 \pm .02$	$.90 \pm .02$	$.98 \pm .01$	$.91 \pm .01$	$.90 \pm .02$	$.94 \pm .02$	$.88 \pm .02$
expert	id. accuracy	-	-	$.91 \pm .02$	-	-	$.87 \pm .05$	-
	quant. accuracy	-	-	$.95 \pm .01$	-	-	$.91 \pm .04$	-

Fig. 5. (left) BAYESIL's true/false positive/negative rate in identification of individual compounds in 50 biological serum samples. (right) The average concentration for correctly identified compounds in the same samples. The error bars show the average difference between BAYESIL and expert values for each compound and the red dots show the detection threshold for each compound.

is not uniform across metabolites. BAYESIL also uses a relative factor in compound detectability; as some compound such as Choline are easy to identify and quantify at low concentrations while for some other compounds such as L-Asparagine, experts use a higher detection threshold.

Given a spectrum of a mixture of compounds (with "true" concentrations $\{\rho_{\mathcal{M}}\}$), BAYESIL returns its *estimates* of these concentrations $\{\widehat{\rho}_{\mathcal{M}}\}$, which might be 0 if that compound is absent. We say a compound is a true positive if both $\widehat{\rho}_{\mathcal{M}}$ and $\rho_{\mathcal{M}}$ are positive – that is, greater than the detection threshold, and a true negative if both $\widehat{\rho}_{\mathcal{M}}$ and $\rho_{\mathcal{M}}$ are less than the threshold; in either case, BAYESIL's prediction is considered correct. BAYESIL's identification accuracy for a given spectrum is the ratio of correct labels (true positives plus true negatives) to the library size. BAYESIL's "quantitative accuracy" describes how often its estimates $\widehat{\rho}_{\mathcal{M}}$ were "close enough" to the true values $\rho_{\mathcal{M}}$; note that simply computing $|\widehat{\rho}_{\mathcal{M}} - \rho_{\mathcal{M}}|$ is not enough as this measure would basically only consider the compounds with high concentrations. We instead use the $\text{median}_{\mathcal{M}} \left(\frac{|\rho_{\mathcal{M}} - \widehat{\rho}_{\mathcal{M}}|}{\max(\widehat{\rho}_{\mathcal{M}}, \rho_{\mathcal{M}})} \right)$ as a measure of the percentage error in concentrations. Table I reports BAYESIL's identification and quantification accuracies, for each of the tasks listed above. For the biological and synthetic samples, we assume the human expert's assessment is correct.

Figure 5(left) reports the frequency of false/true positives/negatives for individual compounds in 50 serum samples. Figure 5(right) shows the average of $\rho_{\mathcal{M}}$ for correctly identified compounds in 50 serum samples, as reported by NMR experts, the average detection threshold for diffent compounds as well as the average difference $\widehat{\rho}_{\mathcal{M}} - \rho_{\mathcal{M}}$, between BAYESIL and expert's estimate for each compound.

These results on a diverse set of test data suggest that BAYESIL is often within 10% of the expert's estimate, and where the ground truth is known, BAYESIL's metabolic profile is often more accurate than the expert's. BAYESIL's web-page *http://www.bayesil.ca* provides a complete description of all of the studies reported above, showing the fits and the metabolic profiles obtained.

A. Efficiency

On a single 2.8 GHz CPU processor, BAYESIL typically takes less than 5 minutes to profile a serum or CSF spectrum. Over a sustained 24 hour period, BAYESIL should be able to process more than 200 spectra (vs. ~20 spectra/day for a human expert) and accurately identify-&-quantify approximately 50 compounds per spectrum. This corresponds to an output of more than 5000 metabolite measurements a day for a single CPU. It is possible to produce profiles yet faster by reducing the number of particles used in BAYESIL's deconvolution process: reducing this by a factor of 10, reduces the run-time to less than one minute. This speed-up often reduces the identification and quantification accuracy by only 5-10%.

B. Limitations

A key disadvantage of NMR analysis (versus mass spec), in general, is that NMR is relatively insensitive, with a lower limit of detection of $\sim 1 - 5$ μM and a requirement of relatively large sample sizes ($\sim 500 \mu L$). For high-quality spectra with high SNR, such as CSF samples in our study, BAYESIL is often able to achieve its high accuracy within these detection limits.

Key to the high level of performance of BAYESIL is the use of biofluid-specific spectral libraries in its spectral fitting routines (*a.k.a.* targeted profiling). Without these, BAYESIL would be substantially slower and less accurate in both identification and quantification. Therefore users must provide BAYESIL with information about the biofluid being analyzed. Obviously, the mis-identification or mislabelling of a biofluid sample could lead to somewhat poorer results.

This need for prior knowledge about the typical composition of biofluid mixtures has motivated us, and others, to spend considerable efforts to determine the NMR-detectable metabolomes for many biofluids, including human plasma/serum, cerebrospinal fluid, saliva and urine (*e.g.*, [23], [24], [25]), milk and rumen (*e.g.*, [26]), cell extracts (*e.g.*, [27]), cancer cells [28], [29], and many other fluids or extracts.

C. Other systems

There have been a number of software packages recently developed to facilitate spectral profiling and compound identification/quantification by NMR; see Appendix I for details. However, they do not seem to be particularly accurate on real biofluids. Furthermore, many of the packages are not currently publicly available. To date, the largest number of metabolites that has been automatically identified and quantified using publicly available software is 26 compounds, by BATMAN [13]. However, an analysis of this magnitude required several hours of CPU time to process a single spectrum. Furthermore, BATMAN also requires a human expert to perform many of the preliminary steps. We compared BAYESIL to BATMAN on simple computer-generated mixtures, involving 5, 10 and 20 compounds selected from BATMAN's library, as well as a preprocessed human serum sample. For the computer generated spectra, both BATMAN and BAYESIL used identical libraries containing only the relevant compounds. BATMAN achieved $85 - 87\%$ quantification accuracy for the computer-generated mixtures but took 2-9 hours to run, while in all cases BAYESIL achieved $> 98\%$ accuracy in less than

3 minutes. For the serum spectra, BAYESIL used a library of 50 compounds, while BATMAN used a subset of 40 compounds that its library has in common with the serum metabolome. BATMAN took 19 hours to analyze the serum spectrum and identified all the compounds in its library (resulting in 85% identification accuracy) and only 8% quantification accuracy compared to BAYESIL which took 5 minutes to achieve 98% identification and 90% quantification accuracy.

CONCLUSION

NMR is a particularly appealing platform for conducting metabolomic studies on biofluids as it is a rapid, robust, highly reproducible, non-destructive, and fully quantitative technique that requires no prior compound separation or derivatization. The main barrier preventing widespread adoption of NMR-based metabolomics in industrial or clinical applications is the requirement for manual spectral profiling. BAYESIL addresses this critical problem by providing fully automated spectral processing and deconvolution. Furthermore it is able to perform this task on mixtures containing > 50 compounds, with 90% accuracy. We believe that removing the automation barrier will have a significant, positive impact on NMR spectroscopy and NMR-based metabolomics. BAYESIL is freely available for users to perform metabolic profiling of 1D ^{1}H NMR spectra of serum, CSF and other biofluid mxitures (excluding urine) collected at 500 and 600 MHz.

III. MATERIALS AND METHODS

To produce each of the reference spectra for BAYESIL's library, we first prepared stock solutions (1 mM to 100 mM) for each compound in 1 L in volumetric flasks. The metabolites were dissolved in 20 mM NaHPO$_4$ (pH 7.0). These stock solutions were further diluted if necessary to obtain a final stock solution concentration of 1 mM. The final sample for NMR was prepared by transferring 1140 μL to a 1.5 mL Eppendorf tube followed by the addition of 140 μL D$_2$O and 120 μL of the reference standard solution (11.67 mM DSS (disodium-2,2-dimethyl-2-silapentane-5-sulphonate), 20 mM NaHPO$_4$, pH 7.0). After confirming that the pH of the sample was between 6.8 and 7.2 (adjusting the buffer if necessary), we transferred 700 μL to a standard NMR tube for spectral acquisition. All library ^{1}H NMR spectra were collected on both 500 MHz and 600 MHz Inova spectrometers equipped with 5 mm Z-gradient PFG probes. A standard presaturation ^{1}H-NOESY experiment(tnnoesy.c) was acquired at 25°C using the first increment of the presaturation pulse sequence. A 4 s acquisition time, a 100 ms mixing time, a 10 ms recyle delay and a 990 ms saturation delay were chosen. Thirty-two transients were acquired for samples collected at 600 MHz while 128 transients were acquired for all samples collected at 500 MHz. Eight steady state scans were employed and the presaturation pulse power was calibrated to provide a field width no greater than 80 Hz. Both the transmitter offset and the saturation pulse were centered on the water resonance and no suppression gradients were used. After spectral collection, the spectra were checked for quality and then analyzed using a locally developed spectral analysis tool to convert the spectra into a series of XML files. In producing the XML library, most peak clusters were given a default shift-window of 0.025 ppm, with the exception of few compounds such as histidine or citrate that are known to be highly pH-sensitive. Both the synthetic and real biological spectral data were collected in the manner described above except for biological CSF in which 1024 scans were collected to compensate for dilution. For sample preparation, CSF was used as is, while serum was obtained after the blood had clotted for 30 min at 25°C and then passed through pre-rinsed 3000 MWCO Amicon Ultra-0.5 filters to remove remaining proteins. In each case 285 μL of filtrate was obtained and 35 μL of D$_2$O and 30 μL of buffer was added. A total of 350 μL was then transferred to a suitable Sigma tube for NMR data acquisition. In the case of biological CSF, where less than 285 μL was obtainable, the samples were diluted with sufficient H$_2$O.

REFERENCES

[1] **Wishart et al. (2007). HMDB: the human metabolome database.** *Nucleic Acids Res, 35(suppl 1), D521-D526.*

[2] *Gieger, C. et al. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet, 4(11), e1000282.*

[3] *Illig, T. et al. (2010). A genome-wide perspective of genetic variation in human metabolism. Nat Genet, 42(2), 137-141.*

[4] *Keurentjes, J. J. et al. (2006). The genetics of plant metabolism. Nature genetics, 38(7), 842-849.*

[5] *Assfalg, M., Bertini, I., Colangiuli, D., Luchinat, C., Schfer, H., Schtz, B., Spraul, M. (2008). Evidence of different metabolic phenotypes in humans. Proc Natl Acad Sci U S A, 105(5), 1420-1424.*

[6] *Gerszten, R. E., Wang, T. J. (2008). The search for new cardiovascular biomarkers. Nature, 451(7181), 949-952.*

[7] *Sreekumar, A. et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature, 457(7231), 910-914.*

[8] *Wishart, D. (2008). Quantitative metabolomics using NMR. Trends Analyt Chem, 27(3), 228-237.*

[9] *Weljie, A. M., Newton, J., Mercier, P., Carlson, E., Slupsky, C. M. (2006). Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Anal Chem, 78(13), 4430-4442.*

[10] *Tredwell, G. D., Behrends, V., Geier, F. M., Liebeke, M., Bundy, J. G. (2011). Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. Anal Chem, 83(22), 8683-8687.*

[11] *Ravanbakhsh, S. (2009) A stochastic optimization method for partially decomposable problems, with applications to analysis of NMR spectra, M.Sc. Thesis, U of Alberta.*

[12] *Mercier, P., Lewis, M. J., Chang, D., Baker, D., Wishart, D. (2011). Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. J Biomol NMR, 49(3-4), 307-323*

[13] *Hao, J., Astle, W., De Iorio, M., Ebbels, T. M. (2012). BATMANan R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. Bioinformatics, 28(15), 2088-2090.*

[14] *Ravanbakhsh, S., Poczos, B., Greiner, R. (2010). A Cross-Entropy method that optimizes partially decomposable problems: a new way to interpret NMR spectra, Proc Conf AAAI Artif Intell.*

[15] *Brown, D. E., Campbell, T. W., Moore, R. N. (1989). Automated phase correction of FT NMR spectra by baseline optimization. J Magn Reson, 85(1), 15-23.*

[16] *de Brouwer, H. (2009). Evaluation of algorithms for automated phase correction of NMR spectra. J Magn Reson, 201(2), 230-238.*

[17] *Dietrich, W., Rdel, C. H., Neumann, M. (1991). Fast and precise automatic baseline correction of one-and two-dimensional NMR spectra. J Magn Reson, 91(1), 1-11.*

[18] *Fritsch, F. N., Carlson, R. E. (1980). Monotone piecewise cubic interpolation. SIAM J Numer Anal, 17(2), 238-246.*

[19] *Whittaker, E. T. (1922). On a new method of graduation. Proceedings of the Edinburgh Mathematical Society, 41, 63-75.*

[20] *Savitzky, A., Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Anal Chem, 36(8), 1627-1639.*

[21] *Koller, D., Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.*

[22] *Doucet, A. (2001). Sequential monte carlo methods. John Wiley and Sons, Inc.*

[23] *Psychogios et al. (2011). The human serum metabolome. PLoS One, 6(2), e16957.*

[24] *Wishart et al. (2008). The human cerebrospinal fluid metabolome. J Chromatogr B Biomed Sci Appl, 871(2), 164-173.*

[25] *Bouatra et al. (2013). The human urine metabolome. PloS One, 8(9), e73076.*

[26] *Sundekilde, U. K., Larsen, L. B., Bertram, H. C. (2013). NMR-based milk metabolomics. Metabolites, 3(2), 204-222.*

[27] *Dietmair, S., Timmins, N. E., Gray, P. P., Nielsen, L. K., Kromer, J. O. (2010). Towards quantitative metabolomics of mammalian cells: Development of a metabolite extraction protocol. Anal Biochem, 404(2), 155-164.*

[28] *Griffin, J. L., Shockcor, J. P. (2004). Metabolic profiles of cancer cells. Nat Rev Cancer, 4(7), 551-561.*

[29] *Abate-Shen, C., Shen, M. M. (2009). Diagnostics: The prostate-cancer metabolome. Nature, 457(7231), 799-800.*

APPENDIX

I. OTHER NMR-ANALYSIS SOFTWARE TOOLS

Several software packages have been developed to facilitate NMR spectral processing, compound identification and quantification. While some facilitate two dimensional NMR spectral processing and compound identification (*e.g.*, [1], [2], dataChord (One Moon Scientific)), none are completely automated and provide compound quantification. Furthermore, because modern metabolomic NMR studies require the analysis of a tremendous amount of spectral data, the time required to collect 2D NMR spectra makes their implementation prohibitive not only due to their inherent inability to facilitate high-throughput studies but also due to complicating factors such as sample degradation within the spectral acquisition time frame. Thus, more effort has been directed at developing software tools for 1D NMR.

Some tools provide basic processing functionalities [3], [4]. (In fact, BAYESIL uses NMR-GLUE [3] for handling different input/output data formats.)

We focus on software package capable of handling high-throughput 1D metabolomic NMR data. An ideal tool here should have the following features: (1) *fully automated* – in both spectral processing and compound identification and quantification; (2) *flexible and customizable* – capable of analyzing a wide (and extendable) range of different biological fluids; (3) *ubiquitous* – can accommodate input from different NMR vendors, multiple spectrometer frequencies; and of course (4) *accurate*.

Various software packages have made incremental steps toward achieving these goals. Of the 19 that we could identify, only a handful provide some degree of automated identification and/or quantification (1): [1], [5], [6], [12], [8], [10], [12], Juice Screener, Wine Screener and Metabolic Profiler (Bruker Corporation) and Chenomx NMR Suite (Chenomx Inc.). This task (of spectral deconvolution and metabolite profiling) has been tackled with a variety of algorithmic approaches – including simple text file matching, binning [5], principal component analysis and non-negative matrix factorization [1], [6], combinations of simulated annealing and gradient descend ([7], Chenomx NMR Suite (Chenomx Inc.)), cross entropy method [8] and Monte Carlo techniques [10]. However only a few of these softwares provide automated spectral processing ([11],Juice Screener and Wine Screener).

In terms of flexibility and customizability (2), some software packages do utilize large data-sets (HMDB [14], BMRB [15] or MMCD [16]) but they still require the user to select a subset of the compounds, and/or do not provide quantification [1], [9], [2]. Others are specialized to particular mixtures ([11], Wine Screener, Juice Screener and Vantera (LipoScience Inc.)) and none can accurately quantify complex mixtures (with > 50 compounds). Moreover, many of these software packages are specific to a particular instrument (*e.g.*, [10], [14], [12], [13], Wine Screener, Juice Screener, dataChord and Vantera).

It is often difficult to access accuracy (4), as the descriptions of many software tools do not provide any assessment (*e.g.*, Juice Screener, Wine Screener and Metabolic Profiler (Bruker Corporation), [1], [11]) and many systems have been assessed merely on very simple mixtures (*e.g.*, [2], [9], [10],lipoprofiler/Vantera (LipoScience Inc.)) or simple spike-in experiments [5], [12], [13].

To summarize, BAYESIL is the only 1D ^{1}H NMR interpretation system that is completely automated (both preprocessing and deconvolution) for a wide range of complex mixtures (*i.e.*, all mammalian biofluids covered by its current library; see *www.bayesil.ca*), involving > 60 compounds. Moreover, it is efficient, general, accurate and publicly available.

APPENDIX

II. BAYESIL'S LOSS FUNCTION

In the manuscript, we described a simple sum of squared error loss function (eq[2]) to explain the basic ideas behind BAYESIL's spectral deconvolution. In practice BAYESIL uses a more complicated loss function that also penalizes the derivatives of the difference between the given spectrum $\mathsf{s}(\cdot)$ and its reconstruction $\widehat{\mathsf{s}}(\cdot\,;\boldsymbol{\rho},\boldsymbol{\delta})$:

$$\ell_{\mathcal{X}}\left(\mathsf{s}(\cdot),\widehat{\mathsf{s}}(\cdot\,;\boldsymbol{\rho},\boldsymbol{\delta})\right) = \sum_{c\in\{0,1,2,3\}} \gamma_c \int_{\mathcal{X}} \left(\frac{\partial^c}{\partial x^c}(\mathsf{s}(x)-\widehat{\mathsf{s}}(x\,;\boldsymbol{\rho},\boldsymbol{\delta}))\right)^2 \mathrm{d}x$$

where the integral for $c = 0$ corresponds to sum of squared errors (eq[2]), and $c \geq 1$ enforce the smoothness of the difference between $\mathsf{s}(\cdot)$ and $\widehat{\mathsf{s}}(\cdot)$. Here, the scalars γ_c weight the relative importance of these terms. Similar to the sum of square errors, this loss function also decomposes over the regions $\mathcal{X}_I$, allowing for the same kind of factor-graph representation.

APPENDIX

III. DETAILS ABOUT THE CONSTRUCTION OF NMR SPECTRAL REGIONS

The construction of the spectral regions is based on the observation that the influence of each peak (and hence of each spectral cluster) is significant over only a relatively small region of the spectrum. To estimate this "region of effect" for each cluster, BAYESIL first obtains an upper-bound $\overline{\rho}_{\mathcal{M}} \geq \rho_{\mathcal{M}}$ on the concentration of each compound. This upper-bound is also used in performing approximate inference (see Appendix IV). For each compound $\mathcal{M}$, this is the minimum of the upper-bounds obtained using each of its clusters $c \in \mathcal{M}$: $\overline{\rho}_{\mathcal{M}} = \min_{c \in \mathcal{M}} \overline{\rho}_{\mathcal{C}}$. The upper-bound from each cluster ($\overline{\rho}_{\mathcal{C}}$) is obtained by progressively shifting the signature for that cluster under the spectrum and finding the maximum value that it can take, assuming all the other compounds and clusters are absent:

$$\overline{\rho}_{\mathcal{C}} \;=\; \max_{\delta_{\mathcal{C}}} \min_x \frac{\mathsf{s}(x)}{\mathsf{s}_{\mathcal{C}}(x;1,\delta_{\mathcal{C}})}$$

where $\mathsf{s}_{\mathcal{C}}(.,1,\delta_{\mathcal{C}})$ is the signature of cluster c as defined by the set of all its peaks (see eq[1]) assuming a unit concentration and allowing $\delta_{\mathcal{C}}$ to vary in a small window $[\underline{\delta}_{\mathcal{C}},\overline{\delta}_{\mathcal{C}}]$ around the center defined by the library.

Now that we have an upper-bound on concentrations, using an example from Figure 2 (in the main manuscript) we show why the region of effect for each cluster is bounded. In this figure, the center ($\delta_{L-Isolucine(1)}$) for the first cluster of *L-Isoleucine* can only appear in the interval [0.9130, 0.9380] PPM; as its concentration is at most $\overline{\rho}_{L-Isoleucine} = 95\mu$M, its contribution to any point outside [.8563, .9954] will be five times less than the estimated *noise-level* of this spectrum, where the noise-level is estimated as the standard deviation of the spectrum $\mathsf{s}(\cdot)$ over all of the baseline points. We can therefore identify this L-Isoleucine cluster with the interval [.8563, .9954]. Note this cluster includes 3 peaks. In general, the range of a cluster spans the set of peaks that it contains.

APPENDIX

IV. DETAILS OF BAYESIL'S APPROXIMATE INFERENCE

This appendix provides details of BAYESIL's inference procedure, using the factor-graph representation of the problem.

Recall that ρ_I and δ_I denote the set of shift and concentration values for all the clusters and associated compounds that can appear in the region $\mathcal{X}_I$; we let $\mu_I = [\rho_I, \delta_I]$ denote the set of variables of both both types. Since the loss function $\ell(\cdot,\cdot)$ is additive over domain the $\mathcal{X}$, we can rewrite the optimization of eq[3] in exponential form as:

$$\boldsymbol{\mu}^* \;=\; \arg_{\mu} \max \prod_I \mathsf{f}_I(\boldsymbol{\mu}_I)$$

$$\mathsf{f}_I(\boldsymbol{\mu}_I) = \exp\left\{ -\frac{1}{T}\ell_{\mathcal{X}_I}(\mathsf{s}(\cdot),\widehat{\mathsf{s}}(.;\boldsymbol{\mu}_I)) \right\} \qquad (7)$$

where each factor f_I is basically the exponential of the negative loss function $(-\ell_{\mathcal{X}_I})$ over the region $\mathcal{X}_I$ and T is the temperature. Here a factor node f_I is connected to all of its associated variables $\mu_i \in \mu_I$ in the factor-graph; see Figure 3. In the following, we use $\partial\mu_i = \{f_I \mid \mu_i \in \mu_I\}$ to refer to all the factors that are adjacent to variable μ_i in the factor-graph.

Inference in this factor-graph is challenging as its factors can each depend on a large number of continuous variables. This means that the most basic task of (conditional) *sampling* from a factor is unfeasible and we cannot use Glauber dynamics (*e.g.*, [18]). This is further complicated by multi-modality of the factors, which prevents the use of parametric densities and inference techniques such as Gaussian Belief Propagation [21] or (primal and dual) decomposition methods that require convex sub-problems [17]. BAYESIL uses a non-parametric sequential Monte Carlo method that is closely related to sequential importance sampling and particle filters. The following is a step-by-step explanation of this inference procedure.

BAYESIL models a distribution $\mathbb{P}(\mu_i)$ over individual variables, non-parametrically: as a set of "particles". These particles, $\mu_i[n]$ for $1 \leq n \leq N$, collectively serve to approximate the target distribution. In all the experiments described in the manuscript we use $N = 10,000$ such particles, however using a larger number can increase accuracy at the cost of increased run-time.

For each n, the joint set of particles for all variables, $\mu[n] = [\rho[n], \delta[n]]$, corresponds to a complete spectrum – *i.e.*, $\hat{s}(x\,;\rho[n],\delta[n]) = \sum_{\mathcal{M}} \hat{s}(x\,;\mathcal{M},\rho_{\mathcal{M}}[n],\delta_{\mathcal{M}}[n])$ – which means we can compute its loss (eq[2]). BAYESIL calculates the loss for each region $\mathcal{X}_I$, $f_I(\mu_I[n])$ (eq[7]). Since each variable μ_i appears in many regions $\partial\mu_i$, we can "credit" each assignment $\mu_i[n]$ with the loss over all such regions. This allows us to compute a "weight" for each variable μ_i and for each particle n. BAYESIL then uses these weighted sets of particles to produce a new distribution for the variable μ_i, one that prefers values that have less loss. BAYESIL then iterates, using this new distribution, until convergence.

More specifically, BAYESIL first assigns each of the variables μ_i to an initial distribution of values $\mathbb{P}^{(0)}(\mu_i)$ – *e.g.*, $\delta_{L-Isoleucine(1)}$ is drawn uniformly from its chemical shift range [0.9130, 0.9380] PPM, and $\rho_{L-Isoleucine(1)}$ is drawn uniformly from its range [0, 95] μM (see Appendix III for the procedure to estimate the upper bound $\overline{\rho}_{L-Isoleucine(1)} = 95\mu$M). It then iterates $t = 0, 1, 2, \ldots$ over the following four steps:

Step1: It draws $N =$10,000 particles from each $\mathbb{P}^{(t)}(\mu_i)$ independently, producing 10,000 complete assignment to these variables $\mu^{(t)}[n] = [\rho^{(t)}[n], \delta^{(t)}[n]]$ (for $n = 1, 2, \ldots,$10,000), from its current product distribution.

Step2: For each joint particle $\mu^{(t)}[n]$: For each factor I, BAYESIL computes the loss associated with this region, $\ell_I(\mu_I[n]^{(t)})$, and then $f_I(\mu_I[n]^{(t)})$ using eq[7].

Step3: Recall each variable μ_i belongs to a set of regions (each corresponding to a factor), $\partial\mu_i$. BAYESIL then implicitly identifies $\mu_i[n]^{(t)}$ with the loss associated with all corresponding regions. This is achieved by defining a weight

$$\omega(\mu_i[n]^{(t)}) \propto \frac{\prod_{f_I \in \partial\mu_i} f_I(\mu_I[n]^{(t)})}{\mathbb{P}^{(t)}(\mu_i[n]^{(t)})} \qquad (8)$$

where $\mathbb{P}^{(t)}(\cdot)$ is the distribution used for sampling – *i.e.*, importance sampling weight.

Step4: Finally, BAYESIL produces a "new" distribution for each variable μ_i, using kernel density estimation (KDE) over a weighted set of the N particles $\mu[n]^{(t)}$ $1 \leq n \leq N$ to represent the marginal distribution $\mathbb{P}^{(t+1)}(\cdot)$ over each variable $\mu_i \in \mu$:

$$\mathbb{P}^{(t+1)}(\mu_i) \propto \sum_{n=1}^{N} \omega(\mu_i[n]^{(t)})\, \mathsf{k}\left(\frac{\mu_i - \mu_i[n]^{(t)}}{h}\right)$$

where $\mathsf{k}(\cdot)$ is a kernel function (*e.g.*, a Gaussian) and the kernel bandwidth h is estimated from the data (*e.g.*, [19]).

BAYESIL then checks for convergence; if convergence occurs, it returns the mode of individual distribution as its approximation to the MAP assignment. If not, it returns to Step1.

Note the temperature parameter T (used in eq[7] which appears in eq[8]) is gradually reduced from a large value towards zero per iteration. Figure 4 is basically showing the evolution of KDEs over the chemical shift variables ($\mathbb{P}^{(t)}(\delta_{\mathcal{C}})$ for $t \in \{1,\ldots,6\}$) over 6 iterations of spectral deconvolution.

In practice, BAYESIL ignores the importance sampling weights. This biases $\mathbb{P}^{(t+1)}(\mu_i)$ towards $\mathbb{P}^{(t)}(\mu_i)$, but significantly reduces the variance and the computation time. Since we are interested in the mode (rather than the marginals) of $\mathbb{P}(\mu_i)$, this trade-off is favourable, as the bias of the previous estimate is mostly towards more probable assignments.

APPENDIX

V. LIST OF NMR-DETECTABLE COMPOUNDS IN SERUM AND CSF

Our serum library includes the following metabolites plus DSS: 1-Methylhistidine, 2-Hydroxybutyric acid, Acetic acid, Betaine, Acetoacetic acid, L-Carnitine, Creatine, Citric acid, Choline, Ethanol, D-Glucose, Glycine, Glycerol, Formic acid, L-Glutamic acid, Hypoxanthine, L-Tyrosine, L-Phenylalanine, L-Alanine, L-Proline, L-Threonine, L-Asparagine, L-Isoleucine, L-Histidine, L-Lysine, L-Serine, L-Lactic acid, L-Aspartic acid, L-Cystine, Ornithine, Pyruvic acid, Succinic acid, Urea, 3-Hydroxybutyric acid, L-Arginine, Creatinine, L-Cysteine, L-Glutamine, L-Leucine, Malonic acid, L-Methionine, Isopropyl alcohol, L-Valine, L-Tryptophan, Acetone, Isobutyric acid, Methanol, Propylene glycol, Dimethyl sulfone.

Our CSF library includes the following compounds plus DSS: 2-Hydroxybutyrate, 2-Oxoisovalerate, 3-Hydroxyisobutyrate, Acetate, Ascorbic acid, Acetoacetate, Creatine, Dimethylamine, Citrate, Choline, Glucose, Glycerol, Formate, Glutamate, Tyrosine, Phenylalanine, Alanine, Threonine, Mannose, Isoleucine, Histidine, Lysine, Serine, Lactate, 2-Oxoglutarate, myo-Inositol, Oxalacetate, Pyruvate, Succinate, Pyroglutamate, Xanthine, Urea, 3-Hydroxybutyrate, 2-Hydroxyisovalerate, Creatinine, Glutamine, Fructose, Leucine, Methionine, 3-Hydroxyisovalerate, Isopropanol, Valine, Tryptophan, Acetone, Methanol, Propylene glycol, 1,5-Anhydrosorbitol, Dimethylsulfone.

REFERENCES

[1] Robinette, S. L., Zhang, F., Bruschweiler-Li, L., Bruschweiler, R. (2008). Web server based complex mixture analysis by NMR. *Anal Chem, 80(10), 3606-3611.*

[2] Xia, J., Bjorndahl, T. C., Tang, P., Wishart, D. (2008). MetaboMinersemi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC bioinformatics, 9(1), 507.*

[3] Helmus, J. J., Jaroniec, C. P. (2013) Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *Journal of biomolecular NMR, 55(4), 355-367.*

[4] Delaglio, F. et al. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR, 6(3), 277-293.*

[5] Zheng, C., Zhang, S., Ragg, S., Raftery, D., Vitek, O. (2011). Identification and quantification of metabolites in 1H NMR spectra by Bayesian model selection. *Bioinformatics, 27(12), 1637-1644.*

[6] Zhao, Q., Stoyanova, R., Du, S., Sajda, P., Brown, T. R. (2006). HiResa tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics, 22(20), 2562-2564.*

[7] Mercier, P., Lewis, M. J., Chang, D., Baker, D., Wishart, D. (2011). Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra. *J Biomol NMR, 49(3-4), 307-323*

[8] Ravanbakhsh, S., Poczos, B., Greiner, R. (2010). A Cross-Entropy method that optimizes partially decomposable problems: a new way to interpret NMR spectra. *Proc Conf AAAI Artif Intell*

[9] Tulpan, D., Leger, S., Belliveau, L., Culf, A., Cuperlovic-Culf, M. (2011). *MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. BMC bioinformatics, 12(1), 400.*

[10] Hao, J., Astle, W., De Iorio, M., Ebbels, T. M. (2012). *BATMANan R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. Bioinformatics, 28(15), 2088-2090.*

[11] Provencher, S. W. (1993). *Estimation of metabolite concentrations from localized in vivo proton NMR spectra.Magnetic Resonance in Medicine, 30(6), 672-679.*

[12] Mihaleva, V. V. et al. (2014). *Automated quantum mechanical total line shape fitting model for quantitative NMR-based profiling of human serum metabolites. Anal Bioanal Chem, 406(13), 3091-3102.*

[13] Crockford, D. J., Keun, H. C., Smith, L. M., Holmes, E., Nicholson, J. K. (2005). *Curve-fitting method for direct quantitation of compounds in complex biological mixtures using 1H NMR: application in metabonomic toxicology studies. Anal Chem, 77(14), 4556-4562.*

[14] Wishart et al. (2007). *HMDB: the human metabolome database. Nucleic Acids Res, 35(suppl 1), D521-D526.*

[15] Eldon L. Ulrich et al. (2008) *BioMagResBank, Nucleic Acids Res, 36, D402-D408*

[16] Q. Cui et al. (2008) *Metabolite identification via the Madison Metabolomics Consortium Database", Nat Biotechnol, 26,162*

[17] Boyd, S. P., Vandenberghe, L. (2004). *Convex optimization. Cambridge university press.*

[18] Mezard, M., Montanari, A. (2009). *Information, physics, and computation. Oxford University Press.*

[19] Silverman, B. W. (1986). *Density estimation for statistics and data analysis (Vol. 26). CRC press.*