THE JOURNAL OF PHYSICAL CHEMISTRY B

pubs.acs.org/JPCB

Article

Bayesian Quantification for Coherent Anti-Stokes Raman Scattering Spectroscopy

Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

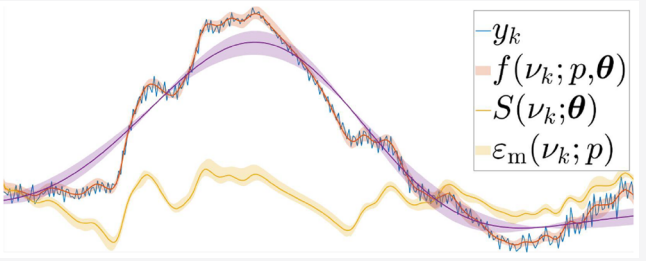
Teemu Härkönen,* Lassi Roininen, Matthew T. Moores, and Erik M. Vartiainen

Cite This: *J. Phys. Chem. B* 2020, 124, 7005−7012

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We propose a Bayesian statistical model for analyzing coherent anti-Stokes Raman scattering (CARS) spectra. Our quantitative analysis includes statistical estimation of constituent line-shape parameters, the underlying Raman signal, the error-corrected CARS spectrum, and the measured CARS spectrum. As such, this work enables extensive uncertainty quantification in the context of CARS spectroscopy. Furthermore, we present an unsupervised method for improving spectral resolution of Raman-like spectra requiring little to no *a priori* information. Finally, the recently proposed wavelet prism method for correcting the experimental artifacts in CARS is enhanced by using interpolation techniques for wavelets. The method is validated using CARS spectra of adenosine mono-, di-, and triphosphate in water, as well as equimolar aqueous solutions of D-fructose, D-glucose, and their disaccharide combination sucrose.



INTRODUCTION

Coherent anti-Stokes Raman scattering (CARS) spectroscopy offers a unique microscopic tool in biophysics, biology, and materials research.[1−14] In addition to being ideally suited for qualitatively label-free microscopy,[2,3,6] the multiplex approach of CARS can also provide complete (position-dependent) vibrational spectra. In principle, this would allow a quantitative, local analysis of chemical composition.[1,11,13,15−18] However, a CARS measurement does not directly provide any quantitative information. Sophisticated analytical methods are therefore required in order to extract this information from the spectroscopic measurements.

An observed CARS spectrum arises from a coherent addition of both resonant contributions from different vibrational modes and a constant, nonresonant (NR) background contribution. This results in a complex line shape, where the positions, amplitudes, and line widths of each vibrational mode are generally hidden. This is particularly true for condensed-phase samples, where the vibrational spectra are highly congested with strongly overlapping vibrational modes.[19] At a minimum, quantitative analysis requires extracting the Raman line shapes from CARS spectra. This can be done by using a suitable phase retrieval method[19,20] on the normalized CARS spectrum. However, the technology is still limited in terms of comparable and quantitative analysis methods, which remain active and ongoing topics of research.[18,19,21,22] Moreover, the analysis is complicated by experimental errors encountered in obtaining a normalized CARS line-shape spectrum, which leads to an erroneous, nonadditive, and nonconstant background component to the NR background due to the reference CARS intensity not arising from a purely nonresonant third-order susceptibility or due to broadband laser behavior inside the sample.[21,22] If it remains uncorrected, this artifact in the NR background can prevent any quantitative information from being obtained from a CARS measurement. Recently, a procedure based on the wavelet prism decomposition algorithm was proposed to address this issue.[22]

Sequential Monte Carlo (SMC) methods have been successfully applied in a wide variety of contexts, including motion tracking,[23,24] satellite image analysis,[25] medical applications,[26] and geophysics.[27] In spectroscopy, Bayesian methods such as SMC have recently been gaining significant attention from the research community. Bayesian statistical inference has been applied to electrochemical impedance,[28] double electron−electron resonance,[29] time-resolved analysis of gamma-ray bursts,[30] and the estimation of elastic and

Received: May 15, 2020
Revised: July 13, 2020
Published: July 16, 2020

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012

crystallographic features by resonance ultrasound spectroscopy[31] to name a few. In particular, a hierarchical Bayesian approach combining modeling of individual line shapes with a continuous background model, with estimation done via SMC methods, has been introduced for Raman spectroscopy.[32]

The contributions of this study are 3-fold. We introduce a method for correcting experimental artifacts in raw CARS measurements, extending further the existing method based on wavelet prism decomposition.[22] Second, we propose a line-narrowing method with improved properties compared to the line shape optimized maximum entropy linear prediction (LOMEP) method.[33,34] Our method utilizes linear prediction, as in LOMEP, but in contrast circumvents the need of assuming a single *a priori* common line shape for all spectral lines. This constitutes a major improvement over the LOMEP method. Third and foremost, Bayesian inference is introduced to CARS spectrum analysis, extending previously available analysis methods. We formulate a Bayesian inference model that is capable of estimating predictive distributions of the underlying Raman signal, the error-corrected CARS spectrum, and the measurement CARS spectrum. This is enabled by parametric modeling of Voigt line shapes, along with a continuous, wavelet-based model for experimental artifacts.

In what follows, we introduce the Bayesian statistical model for CARS. The Raman signal of the CARS spectrum is modeled using a linear combination of Voigt line shapes. Using the Hilbert transform, we construct the modulus of the resonant part of the CARS spectrum. A nonresonant part, estimated from the data,[22] is added to obtain an error-free CARS spectrum, which is finally modulated with a slowly varying error function. Next, we describe the numerical algorithms used for statistical inference and line narrowing, along with our Bayesian prior distributions. We then present experimental details along with obtained results for the means of the constituent line shapes and the predictive intervals for the resonant Raman signal, modulating error function, error-corrected CARS spectrum, and the measurement CARS spectrum. Lastly, the key aspects of the study are briefly remarked upon, thereby concluding the paper.

METHODS

Statistical Model. We model CARS spectral measurements with an additive error model given as

$$y_k := y(\nu_k) = f(\nu_k; p, \boldsymbol{\theta}) + \epsilon(\nu_k) \tag{1}$$

where y_k denotes a measurement that has been discretized with spectral sampling resolution $h > 0$ at a wavenumber location $\nu_k = kh$ with $k \in \mathbb{Z}_+$, $f(\nu_k; p, \boldsymbol{\theta})$ is the CARS spectrum model with parameter p controlling the baseline and parameters $\boldsymbol{\theta}$ for the Voigt line shape, and with measurement error $\epsilon(\nu_k) \sim \mathcal{N}(0, \sigma_\epsilon^2)$ with known variance. For the spectrum, we use a parameter-wise separable model

$$f(\nu; p, \boldsymbol{\theta}) = \varepsilon_m(\nu; p) S(\nu; \boldsymbol{\theta}) \tag{2}$$

where p is the interpolated discrete wavelet transform (DWT) detail level, $\varepsilon_m(\nu; p)$ is the modulating error function, and $S(\nu; \boldsymbol{\theta})$ is the error-corrected CARS signal, similar to the representation used in ref 22. The signal S can further be represented as

$$\begin{aligned} S(\nu; \boldsymbol{\theta}) &= \left|\chi^{(3)}_{NR}(\nu) + \chi^{(3)}_{R}(\nu; \boldsymbol{\theta})\right|^2 \\ &= \left|\exp\left(\frac{A_J(\nu)}{2}\right) + (iV_N(\nu; \boldsymbol{\theta}) - \mathcal{H}\{V_N(\nu; \boldsymbol{\theta})\})\right|^2, \end{aligned} \tag{3}$$

where the exponential part corresponds to the non-Raman part with A_J practically constant (see ref 22 for details), $\mathcal{H}$ is the Hilbert transform, and

$$\begin{aligned} V_N(\nu; \boldsymbol{\theta}) &= \sum_{n=1}^{N} a_n V(\nu - \nu_n; \sigma_n, \gamma_n) \\ &= \sum_{n=1}^{N} a_n L(\nu - \nu_n; \gamma_n) * G(\nu - \nu_n; \sigma_n) \\ &= \sum_{n=1}^{N} a_n \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\nu - \nu_n)^2}{2\sigma_n^2}\right) \\ &\quad * \frac{1}{\pi\gamma_n} \frac{\gamma_n^2}{(\nu - \nu_n)^2 + \gamma_n^2}, \end{aligned} \tag{4}$$

where * denotes convolution. N stands for the number of line shapes, with each line shape having $\boldsymbol{\theta}_n := (a_n, \nu_n, \sigma_n, \gamma_n)^T$ parameters standing for the amplitude, location, scale of the Gaussian shape, and scale of the Lorentzian shape, respectively. Thus, we have $4N$ parameters in total for our model of $S(\nu; \boldsymbol{\theta})$.

Instead of the wavelet prism method,[22] we model the modulating error function as

$$\log(\varepsilon_m(\nu; p)) = \sum_{j=\lceil p+1 \rceil}^{J} D_j(\nu) + (1 - \beta) D_{\lceil p \rceil}(\nu) \tag{5}$$

where $p \in [1, J]$, and $\beta = p - \lfloor p \rfloor$, i.e., as an interpolation between the discrete wavelet reconstruction levels D_j to have a continuous model for the background in contrast to the wavelet prism method where only the discrete wavelet reconstruction levels are used. With the above, we can have an unnormalized posterior formulated as

$$\pi(p, \boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\mathbf{y} \mid p, \boldsymbol{\theta}) \pi_0(p, \boldsymbol{\theta}) \tag{6}$$

where $\mathbf{y} := (y_1, \cdots, y_K)^T \in \mathbb{R}^K$ is the vector of observations given via eq 1, $\boldsymbol{\theta} \in \mathbb{R}_+^{4N}$ is the parameter vector $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^T$ for the N Voigt peaks, $\mathcal{L}(\mathbf{y} \mid p, \boldsymbol{\theta})$ represents the likelihood distribution of the forward model, and $\pi_0(p, \boldsymbol{\theta})$ denotes prior distributions for some or all of the model parameters p and $\boldsymbol{\theta}$. As such, the total number of parameters in the model is $4N + 1$. The solution of (6) is unavailable in closed form, but following ref 32, we can use Monte Carlo methods to obtain samples from this distribution, as described in the following section.

Sequential Monte Carlo. Sequential Monte Carlo (SMC) methods, also known as particle filtering and smoothing, are widely used in statistical signal processing.[35] These algorithms provide a general procedure for sampling from Bayesian posterior distributions.[36,37] SMC methods utilize a collection of weighted particles, initialized from a prior distribution, which are ultimately transformed to represent a posterior distribution under investigation. The methodology used in this study is similar to the one used in ref 22, where they use

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012

sequential likelihood tempering[37] to fit a model of surface-enhanced Raman spectra to measurements.

Assuming additive Gaussian measurement errors $\epsilon(\nu_k)$ as in eq 1, the likelihood of the model $f(\nu_k;\ p,\ \boldsymbol{\theta})$ fitting measurement data $\mathbf{y}$ can be formulated as

$$\mathcal{L}(\mathbf{y} \mid p, \boldsymbol{\theta}) \sim \prod_{k=1}^{K} \mathcal{N}(y_k; f(\nu_k;\ p,\ \boldsymbol{\theta}),\ \sigma_\epsilon^2) \tag{7}$$

and the posterior distribution for step t of the sequential likelihood tempering is given by

$$\pi^{(t)}(p, \boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\mathbf{y} \mid p, \boldsymbol{\theta})^{\kappa^{(t)}} \pi_0(p, \boldsymbol{\theta}) \tag{8}$$

where the superscript (t) denotes the iteration or "time" step of the algorithm and $\kappa^{(t)}$, $\kappa^{(t-1)} < \kappa^{(t)} < \kappa^{(t+1)} < \cdots \leq 1$ with $\kappa^{(0)} = 0$, being a parameter controlling the degree of tempering of the likelihood, with the initial state being equal to the prior distribution while increasingly tempering the total likelihood toward the complete Bayes' theorem. The tempering parameter $\kappa^{(t)}$ can be defined simply as an strictly increasing sequence so that $\kappa^{(t)} \in [0,\ 1]$ or as done in ref 32, the parameter can be determined adaptively according to a given learning rate η such that the relative reduction in the ESS between iterations is approximately η.

Using Q particles, with Q being the number of parameter values used to approximate the posterior distribution, individual weights of each particle at initial step $t = 0$ are set as equally important, so $w_q^{(0)} = 1/Q$. The weights are then updated at each step t according to

$$w_q^{(t)} \propto \frac{\mathcal{L}(\mathbf{y}|p, \boldsymbol{\theta})^{\kappa^{(t)}}}{\mathcal{L}(\mathbf{y}|p, \boldsymbol{\theta})^{\kappa^{(t-1)}}} w_q^{(t-1)} \tag{9}$$

and then normalized so that $\sum_{q=1}^{Q} w_q^{(t)} = 1$. However, updating the particle weights gradually impoverishes the sample distribution. This degradation is measured by the effective sample size (ESS)

$$Q_{\mathrm{ESS}}^{(t)} = \frac{1}{\sum_{q=1}^{Q} (w_q^{(t)})^2} \tag{10}$$

To counteract this, the particles are resampled according to a chosen resampling algorithm when the ESS has fallen below a set threshold Q_{min}. The particle weights are then reset as $w_q^{(t)} = 1/Q$. Some duplication of the particles is inevitably introduced due to the resampling procedure. To remove these duplicates, each particle is additionally updated using Markov chain Monte Carlo (MCMC) targeting the invariant distribution given by the tempered posterior defined in eq 8 at the current iteration or "time" step t. The pseudocode of our SMC algorithm is presented in Algorithm 1.

Line Narrowing. We employ a line-narrowing method to obtain an initial estimation of peak locations ν_n, amplitudes a_n, and number of line shapes N. This is a preprocessing step for the statistical estimation method described in the previous section. A spectrum with Lorentzian line shapes can be approximately modeled as

Algorithm 1 A sequential Monte Carlo sampler.

Initialize:
- Set $t = 0$ and $\kappa^{(t)} = 0$.
- Draw Q particles from the prior distribution $\pi_0(p, \boldsymbol{\theta})$.
- Set particle weights $w_q = \frac{1}{Q}$.

while $\kappa_t < 1$ **do**
- $t = t + 1$.
- Determine $\kappa^{(t)}$.
- Update particle weights $w_q^{(t)}$ according to (9).
- Compute the effective sample size $Q_{\mathrm{ESS}}^{(t)}$ using (10).
- **if** $Q_{\mathrm{ESS}}^{(t)} < Q_{\mathrm{min}}$ **then**
  - Resample particles according to their weights.
  - Set particle weights $w_q = \frac{1}{Q}$.
- **end if**
- Update the particles with MCMC using the tempered posterior given by (8).
- Update particle weights $w_q^{(t)}$ according to their likelihoods.
- Recompute the effective sample size $Q_{\mathrm{ESS}}^{(t)}$ using (10).

end while

$$\begin{aligned}\tilde{V}_N(\nu_k, \tilde{\boldsymbol{\theta}}) &:= \sum_{n=1}^{N} a_n L(\nu_k;\ \nu_n,\ \gamma_n) \approx \sum_{n=1}^{N} a_n L(\nu_k;\ \nu_n,\ \gamma) \\ &= L(\nu_k;\ 0,\ \gamma) * \sum_{n=1}^{N} a_n \delta(\nu_k - \nu_n)\end{aligned} \tag{11}$$

where $\tilde{V}_N(\nu_k, \tilde{\boldsymbol{\theta}})$ denotes a spectrum measured at location ν_k with parameters $\tilde{\boldsymbol{\theta}} := (a_n,\ \nu_n,\ \gamma_n)^T$, γ is a single, constant parameter for the line width, and $\delta(\nu - \nu_n)$ is the Dirac delta function.

Our starting point is the LOMEP method,[33,34] where the constant γ approximation is used. With suitably chosen γ, we have

$$\mathcal{F}\left\{\sum_{n=1}^{N} a_n \delta(\nu - \nu_n)\right\} = \frac{\mathcal{F}\{\tilde{V}_N(\nu_k, \tilde{\boldsymbol{\theta}}))\}}{\mathcal{F}\{L(\nu_k;\ 0,\ \gamma)\}} =: x_{\mathrm{LP}}(t_k;\ \gamma,\ N_{\mathrm{FIR}}) \tag{12}$$

where $\mathcal{F}$ denotes the Fourier transform, t_k the Fourier domain variable, and $x_{\mathrm{LP}}(t_k;\ \gamma_m,\ N_{\mathrm{FIR}})$ is the linearly predicted time signal. In LOMEP, the linear prediction is done using finite impulse response filtering with filter length $N_{\mathrm{FIR}} - 1$.[33,34] The major limitation of LOMEP is the heuristic choice of γ. Additionally, the q-curve optimization method fails when the number of line shapes N increases. Despite these drawbacks, the potential of the linear prediction scheme is nevertheless attractive for its ability to substantially sharpen the line shapes when it is successful.

As an alternative to the approximation model in eq 11, we propose a linear combination of M similarly constructed convolutions

$$\begin{aligned}\sum_{n=1}^{N} a_n L(\nu;\ \nu_n,\ \gamma_n) &\approx \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} a_n L(\nu;\ \nu_n,\ \gamma_m) \\ &= \frac{1}{M} \sum_{m=1}^{M} \left(L(\nu;\ 0,\ \gamma_m) * \sum_{n=1}^{N} a_n \delta(\nu - \nu_n) \right)\end{aligned} \tag{13}$$

using a set of width parameters γ_m in contrast to fixed γ. Then, the approximation of the Dirac delta functions is

$$\begin{aligned}D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}}) &= \mathcal{F}^{-1}\{x_{\mathrm{LP}}(t_k;\ \gamma_m,\ N_{\mathrm{FIR}})\} \\ &\approx \sum_{n=1}^{N} a_n \delta(\nu_k - \nu_n)\end{aligned} \tag{14}$$

The squared sum of residuals for a single convolution, denoted here by $d(\gamma_m, N_{\mathrm{FIR}})$, can be given as

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012

$$d(\gamma_m, N_{\mathrm{FIR}}) = \|\tilde{V}_N(\nu_k, \tilde{\boldsymbol{\theta}}) - L(\nu_k; 0, \gamma_m) * D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})\|_2^2 \tag{15}$$

We additionally define a constrained squared sum of residuals as

$$d_{\mathrm{C}}(\gamma_m, N_{\mathrm{FIR}}) = \left\|\tilde{V}_N(\nu_k, \tilde{\boldsymbol{\theta}}) - c_{\mathrm{n}} L(\nu_k; 0, \gamma_m) * \mathbf{1}_{D_{\mathrm{A}}>0} D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})\right\|_2^2 \tag{16}$$

where $\mathbf{1}_{D_{\mathrm{A}}>0} = 1$, if $D_{\mathrm{A}} > 0$ and 0 otherwise, and c_{n} is a normalization constant so that the area under the spectrum is conserved:

$$c_{\mathrm{n}} = \frac{\sum_{k=1}^{K} D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})}{\sum_{k=1}^{K} \mathbf{1}_{D_{\mathrm{A}}>0} D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})} \tag{17}$$

With $d_{\mathrm{C}}(\gamma_m, N_{\mathrm{FIR}})$, we truncate any negative parts of $D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})$ and distort the truncated spectrum according to the normalization constant c_{n} depending on how much signal energy is present on the negative parts. By Parseval's theorem, and by using an orthonormal wavelet basis, the energy of a signal $g(\nu)$ can be represented as

$$\int_{-\infty}^{\infty} \left|g(\nu)\right|^2 \mathrm{d}t = \sum_{l=-\infty}^{\infty} \left|a(l)\right|^2 + \sum_{j=0}^{\infty} \sum_{\kappa=-\infty}^{\infty} \left|b_j(\kappa)\right|^2 \tag{18}$$

where a and b are the scaling function and wavelet coefficients obtained using DWT. Given a signal with sharp features, the energy of the signal should be concentrated on the wavelet coefficients b_j and, a measure of this concentration of wavelet coefficient energy (we) can be defined as

$$C_{\mathrm{we}} = \frac{\sum_{j=0}^{\infty} \sum_{\kappa=-\infty}^{\infty} \left|b_j(\kappa)\right|^2}{\sum_{l=-\infty}^{\infty} \left|a(l)\right|^2 + \sum_{j=0}^{\infty} \sum_{\kappa=-\infty}^{\infty} \left|b_j(\kappa)\right|^2} \tag{19}$$

With the above formulations, we propose Algorithm 2: Define a set of width parameters γ_m, for example, inferred from computational chemistry. Similarly, define an upper bound for the impulse response parameter N_{FIR}. Then, compute $D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})$ using linear prediction for all parameter combinations of γ_m and N_{FIR} and residuals d and d_{C} along with the wavelet energy concentrations C_{we}.

Using the filtering criterion $f_{\mathrm{c}} = d + d_{\mathrm{C}}$, narrow down the set of possible solutions by sorting them according to f_{c} and C_{we}. Take a percentage p_{we} of the wavelet energy sorted solutions, including the largest energy concentrations. Similarly, take a percentage $p_{f_{\mathrm{c}}}$ of the filtering criterion sorted solutions, including the smallest filtering criteria. Thus, an intersection of these sets should include solutions with mostly positive and sharp line shapes. Sort this intersection set of size $\tilde{M}$ according to d. Finally, estimate eq 13 by choosing M so that the sum of residuals d_M is minimized:

$$\arg\min_{M \le \tilde{M}} d_M = \arg\min_{M \le \tilde{M}} \left\|\tilde{V}_N(\nu_k, \tilde{\boldsymbol{\theta}}) - \frac{1}{M} \sum_{m=1}^{M} L(\nu; 0, \gamma_m) * D_{\mathrm{A}}(\nu_k, \gamma_m, N_{\mathrm{FIR}})\right\|_2^2 \tag{20}$$

As needed, smooth the obtained line-narrowed spectrum with a smoothing function.

Algorithm 2 Line narrowing algorithm.

Initialize:
  Set γ_m.
  Set N_{FIR}.
for γ_m **do**
  for N_{FIR} **do**
    Apply linear prediction using γ_m and N_{FIR}.
    Compute d, d_{C}, and C_{we}.
  end for
end for
Construct the solution:
  Filter out a set of possible solutions according to f_{c} and C_{we}.
  Sort the possible solutions according to d.
  Compute d_M and choose the M solutions which minimize d_M.
Smoothing:
  Convolute the result using an appropriate smoothing kernel.

Priors. We obtained priors by manually correcting for the experimental artifacts modeled by eq 5 and simultaneously applying phase retrieval[19,20,38,39] and computation of the resonant imaginary component of the CARS spectrum until a reasonable Raman signal was observed. The line-narrowing algorithm was applied on the manually estimated Raman signal, producing a line-narrowed spectrum from which individual line shapes could be identified. We follow ref 32 in setting informative priors for the line shape locations ν_k as normal distributions

$$\pi_0(\nu_n) \sim \mathcal{N}(\mu_{\nu_n}, \sigma_{\nu_n}^2) \tag{21}$$

where μ_{ν_n} and $\sigma_{\nu_n}^2$ are estimated for each line shape $V(\nu, \boldsymbol{\theta}_n)$ by numerically integrating perceived individual line shapes in the line-narrowed spectrum to estimate the means μ_{ν_n} and variances $\sigma_{\nu_n}^2$. The line-narrowing algorithm utilizes multiple Lorentzian line shapes with differing scale parameters γ_m, thereby giving access to an informative prior for γ_n. As in ref 32, we set a common prior for each γ_n as a log-normal distribution:

$$\pi_0(\log(\gamma_n)) \sim \mathcal{N}(\mu_{\log(\gamma)}, \sigma_{\log(\gamma)}^2) \tag{22}$$

where the estimates for the mean and variance, $\mu_{\log(\gamma)}$ and $\sigma_{\log(\gamma)}^2$, are obtained from the parameters contained in the intersection set of size $\tilde{M}$. Priors for the Gaussian shape parameters σ_n are obtained by scaling $\pi_0(\log(\gamma_n))$ by $\sqrt{2\log(2)}$. This would correspond to using identical priors for the full-width-at-half-maximum of both the Gaussian and Lorentzian line shapes. For the amplitudes, we can obtain an estimate for the areas straight-forwardly by the same numerical integration used to estimate the priors for the locations, as described above. We set a fairly wide prior for the amplitude by setting them as

$$\pi_0(a_n) \sim \mathcal{N}\left(\mu_{a_n}, \left(\frac{\mu_{a_n}}{4}\right)^2\right) \tag{23}$$

where the mean μ_{a_n} is the numerically integrated area of each line shape. A prior for the background parameter p is set as a uniform prior:

$$\pi_0(p) \sim \mathcal{U}(p_{\min}, p_{\max}) \tag{24}$$

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012

An estimate for the noise level σ_ϵ^2 was also obtained using the line-narrowing algorithm. The algorithm fits a smooth representation of the Raman spectrum to the manually corrected data according to eq 20. This smooth representation of the Raman signal is then transformed to the measurement space by eq 3 and then by eq 2. The resulting residuals between the transformed smooth Raman signal and the measured CARS spectrum were used as an estimate for the noise variance σ_ϵ^2. Detailed descriptions of priors specific for each experimental data set of fructose, glucose, sucrose, and adenosine phosphate can be found in the Supporting Information.

EXPERIMENTAL DETAILS

Samples. The sugar samples used in the multiplex CARS spectroscopy were equimolar aqueous solutions of D-fructose, D-glucose, and their disaccharide combination, sucrose (α-D-glucopyranosyl-(1→2)-β-D-fructofuranoside). For sample preparation, the sugar samples were dissolved in buffer solutions (50 mM HEPES, pH = 7) at equal molar concentrations of 500 mM.[16] The adenosine phosphate sample was an equimolar mixture of AMP, ADP and ATP in water for a total concentration of 500 mM.[19] The adenine ring vibrations[40] are found at identical frequencies for either for AMP, ADP, or ATP around 1350 cm^{-1}. The phosphate vibrations between 900 and 1100 cm^{-1} can be used to discriminate between the different nucleotides.[15] The triphosphate group of ATP shows a strong resonance at 1123 cm^{-1}, whereas the monophosphate resonance of AMP is found at 979 cm^{-1}. For ADP a broadened resonance is found in between at 1100 cm^{-1}.

Multiplex CARS Spectroscopy. All CARS spectra used to validate our methodology were recorded using a multiplex CARS spectrometer, the detailed description of which can be found elsewhere.[1,15] In brief, a 10 ps and an 80 fs mode-locked Ti:sapphire lasers were electronically synchronized and used to provide the narrowband pump/probe and broadband Stokes laser pulses in the multiplex CARS process. The center wavelengths of the pump/probe and Stokes pulses were 710 nm. The Stokes laser was tunable between 750 and 950 nm. The sugar spectra were probed within a wavenumber range from 700 to 1250 cm^{-1}, and the AMP/ADP/ATP spectrum within a range from 900 to 1700 cm^{-1}. The linear and parallel polarized pump/probe and Stokes beams were made collinear and focused with an achromatic lens into a tandem cuvette. The latter could be translated perpendicular to the optical axis to perform measurements in either of its two compartments, providing a multiplex CARS spectrum of the sample and of a nonresonant reference under near-identical experimental conditions. Typical average powers used at the sample were 95 mW (75 mW, in case of AMT/ADP/ATP) and 25 mW (105 mW) for the pump/probe and Stokes laser, respectively. The anti-Stokes signal was collected and collimated by a second achromatic lens in the forward-scattering geometry, spectrally filtered by short-pass and notch filters, and focused into a spectrometer equipped with a CCD camera. The acquisition time per CARS spectrum was 200 ms for sugar spectra and 800 ms for the AMP/ADP/ATP spectrum.

Computational Details. The SMC algorithm was computed using $Q = 2000$ particles with the resampling threshold set to $Q_{min} = 1000$ and the learning parameter set as $\eta = 0.9$. Resampling was done, as in ref 32, via residual resampling.[41] Target MCMC acceptance rate was set to 0.23 and the number of MCMC updates at each iteration was 200. An AMD Ryzen 3950X processor was used with 27 CPU threads utilized, with the SMC estimation taking 580, 522, 413, and 688 s to produce the final posterior estimate of the parameters for the fructose, glucose, sucrose, and phosphate samples, respectively. For modeling the modulating error function $\varepsilon_m(\nu; p)$, symlet 34 basis functions were used.

The line-narrowing algorithm was run with $\gamma_m \in [1, 35]$ linearly spaced using 33 points. The maximum number of measurement points N_{FIR} used was 150, meaning that $N_{FIR} = \{1, ..., 150\}$. The length of the extrapolated signal[33,34] was set to equal the number of measurement points in each spectrum. The percentages p_{we} and p_{f_c} were set as 50% and 2.5% respectively. To ensure that $\tilde{M} > 0$, p_{f_c} was incrementally increased by 2.5% until a minimum intersection set size $\tilde{M} \geq 50$ was achieved. For computation of the wavelet energy concentration C_{we} symlet, eight basis functions were used.

RESULTS AND DISCUSSION

In what follows, 95% predictive intervals for the forward model $f(\nu_k; p, \boldsymbol{\theta})$, the modulating error function $\varepsilon_m(\nu; p)$, and the error-corrected spectra $S(\nu; \boldsymbol{\theta})$ are presented in Figures 1a, 2a, 3a, and 4a for the experimental spectra of fructose, glucose, sucrose, and phosphate, respectively. Similarly, in Figures 1b, 2b, 3b, and 4b the 95% predictive interval for the Raman signal represented by $V_N(\nu, \boldsymbol{\theta})$ is presented along with the means for each constituent line shape $V(\nu, \boldsymbol{\theta}_n)$. To illustrate how the priors were estimated, the manually corrected Raman signal

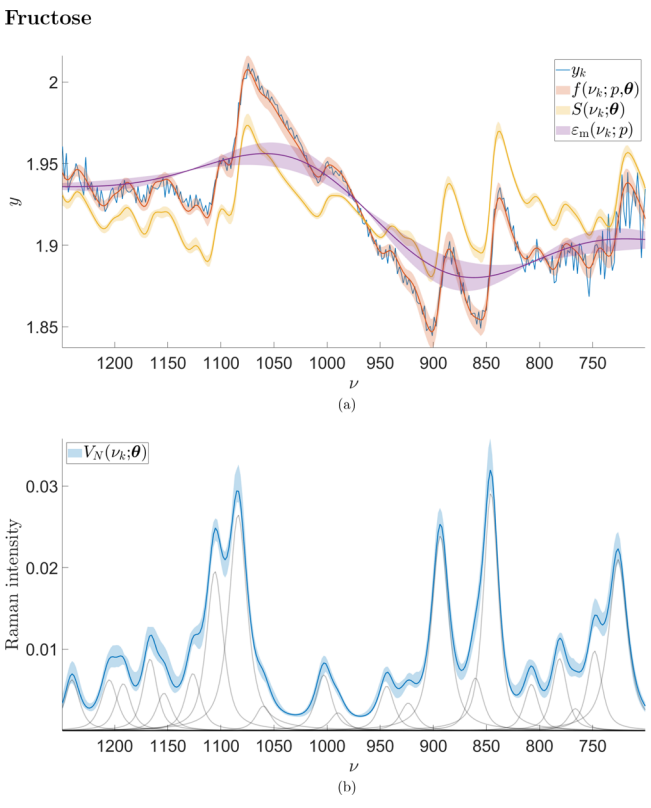


Figure 1. (a) Obtained 95% predictive intervals for y_k, f, S, and ε_m shown in blue, red, yellow, and purple respectively for a CARS measurement of a fructose sample. (b) Obtained 95% predictive intervals for $V_N(\nu_k; \boldsymbol{\theta})$ and means of each individual line shape $V(\nu_k; \boldsymbol{\theta}_n)$ for the fructose sample.

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012

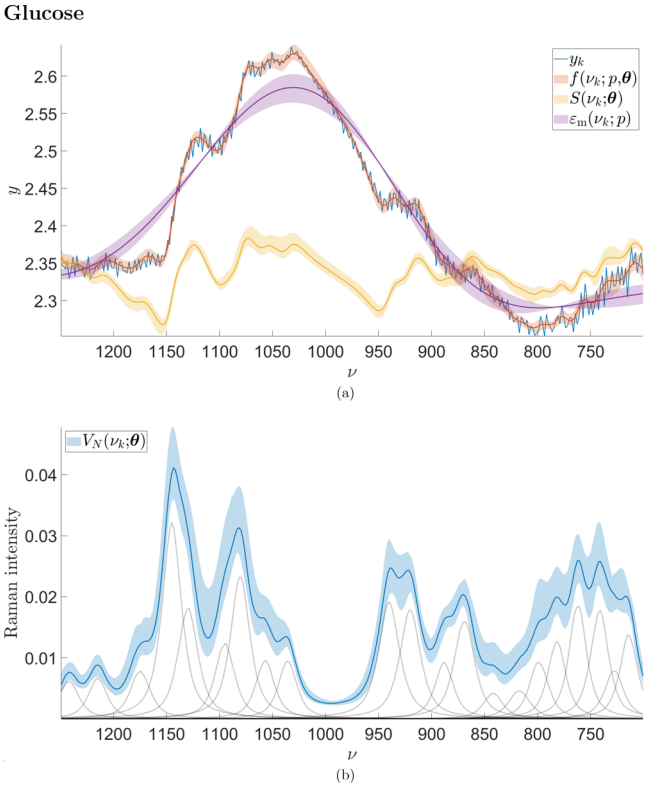


Figure 2. (a) Obtained 95% predictive intervals for y_k, f, S, and ε_m shown in blue, red, yellow, and purple respectively for a CARS measurement of a glucose sample. (b) Obtained 95% predictive intervals for $V_N(\nu_k; \boldsymbol{\theta})$ and means of each individual line shape $V(\nu_k; \boldsymbol{\theta}_n)$ for the glucose sample.

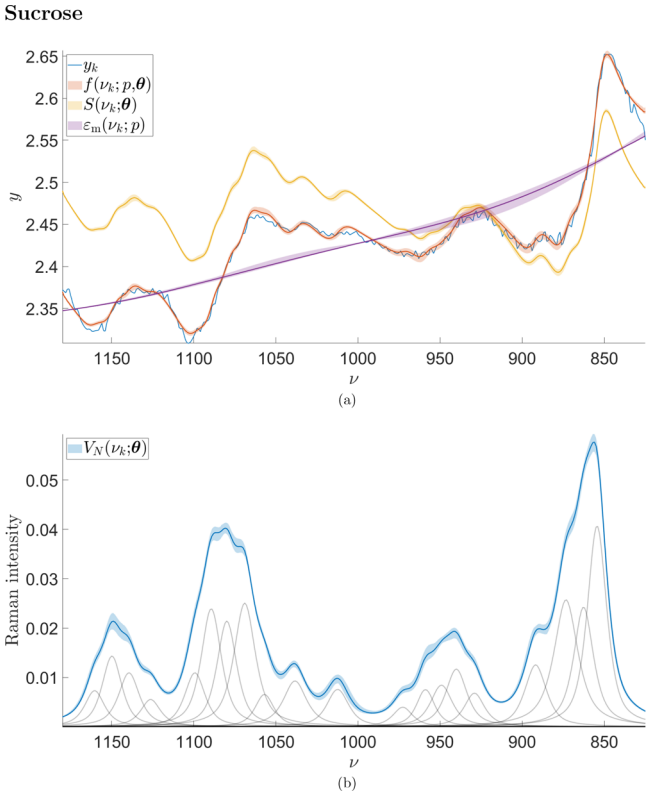


Figure 3. (a) Obtained 95% predictive intervals for y_k, f, S, and ε_m shown in blue, red, yellow, and purple respectively for a CARS measurement of a sucrose sample. Some discrepancies between y_k and f can be seen around the boundaries. These areas of the data should be ignored in the optimization. (b) Obtained 95% predictive intervals for $V_N(\nu_k; \boldsymbol{\theta})$ and means of each individual line shape $V(\nu_k; \boldsymbol{\theta}_n)$ for the sucrose sample.

and the result obtained via the proposed line narrowing method are shown in Figure 5. Additionally, the obtained posterior distributions for $\boldsymbol{\theta}$, alongside their respective prior distributions, are presented in the Supporting Information.

The inference model proposed here was found to adequately model the CARS measurements along with perceived noise levels in the spectra. For future work, it would be interesting to include heteroscedasticity in the model instead of assuming a constant measurement error variance. Comparing the estimated predictive intervals of the obtained Raman signal showed clear correspondence to measured Raman intensities of aqueous solutions for fructose and glucose.[42] To validate the potential of the line narrowing method, we considered the number of line shapes identified for the aqueous solution of sucrose to resemble the 18 line shapes reported for solid sucrose.[43] The estimated priors were considered not to restrict the parameter posterior which can be observed in the posterior distributions when seen alongside the respective priors, especially so for the cases of fructose, sucrose, and adenosine phosphate. The obtained Raman signals for fructose, glucose, and sucrose are similar to results obtained ref 22, which further supports the applicability of the methodology presented in this study. As our method is immediately applicable to more complex samples, such as solutions with multiple solutes, applying the method to such samples provides further interesting future work.

Obtaining informative priors can be approached in different ways for chemically known samples, as was done in ref 32, where the authors use results obtained by density functional theory (DFT) software to derive estimates for the location priors and existing studies on structural properties of a known sample such as observed in refs 42 and 43. Naturally, any other forms of information on the underlying line shapes could just as well be used for the prior distributions. Here we have considered estimating the priors purely from the data using a line-narrowing algorithm, requiring minimal *a priori* information on the sample under study. Although this information would clearly be available in this case,[42] there are many potential applications of our method where much less is known about the molecules in question. Additionally, the use of maximum-entropy methods in improving spectral resolution can cause individual line shapes to split.[44,45] In our proposed line-narrowing method, the averaging together of multiple, resolution-enhanced spectra is postulated to possibly lessen the effect of this undesired spectral line splitting.

CONCLUSION

A Bayesian inference model applicable to coherent anti-Stokes Raman spectroscopy is proposed and numerically implemented. This work extends the current methodology of analyzing CARS spectra by introducing Bayesian inference in the field, enabling uncertainty quantification of spectral features. The statistical inference model is able to produce posterior distributions for physically informative parameters, line shape amplitudes, widths, and locations, for each constituent line shape along with predictive distributions for the estimated resonant Raman signal contained in the CARS measurement spectrum, the error-corrected CARS measure-

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012

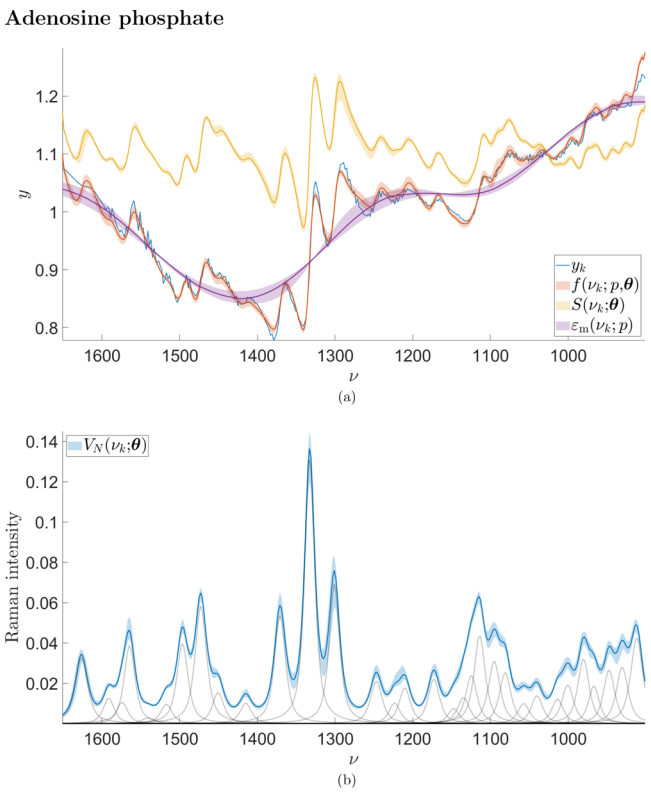


Figure 4. (a) Obtained 95% predictive intervals for y_k, f, S, and ε_m shown in blue, red, yellow, and purple respectively for a CARS measurement of a adenosine phosphate sample. (b) Obtained 95% predictive intervals for $V_N(\nu_k; \boldsymbol{\theta})$ and means of each individual line shape $V(\nu_k; \boldsymbol{\theta}_n)$ for the adenosine phosphate sample.

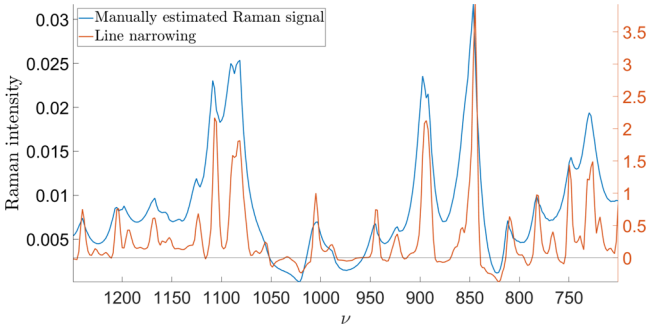


Figure 5. Manually estimated Raman signal, according to the procedure described in the section Priors, of the fructose sample and the line-narrowed Raman signal are shown in blue and red, respectively. The perceivable individual line shapes were numerically integrated to yield informative prior estimates for Voigt line shape parameters.

ments, and the CARS measurement spectrum, as well as extending currently existing methodology for modeling experimental artifacts present in CARS measurements. Additionally, we have developed a line-narrowing algorithm requiring minimal *a priori* information on the underlying line shapes, which is readily applicable to various spectral measurements. We have successfully used this algorithm to obtain informative priors purely from the measurement data for the Bayesian inference model. The applicability of the methods is demonstrated with experimental CARS spectra of sucrose, fructose, glucose, and adenosine phosphate.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.0c04378.

Model parameter priors and obtained posterior distributions of the parameters for each case are available online (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Teemu Härkönen − *LUT School of Engineering Science, LUT University, FI-53851 Lappeenranta, Finland;* orcid.org/0000-0001-5731-6872; Email: teemu.harkonen@lut.fi

Authors

Lassi Roininen − *LUT School of Engineering Science, LUT University, FI-53851 Lappeenranta, Finland*

Matthew T. Moores − *National Institute for Applied Statistics Research Australia, University of Wollongong, Keiraville, NSW 2500, Australia*

Erik M. Vartiainen − *LUT School of Engineering Science, LUT University, FI-53851 Lappeenranta, Finland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.0c04378

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Prof. Heikki Haario for useful discussions and Michiel Müller and Hilde Rinia for providing the experimental data. This work has been funded by the Academy of Finland (Project Numbers 312122, 326341 and 327734). M.T.M. also thanks the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (Project Number CE140100049).

■ REFERENCES

(1) Müller, M.; Schins, J. M. Imaging the Thermodynamic State of Lipid Membranes with Multiplex CARS Microscopy. *J. Phys. Chem. B* **2002**, *106*, 3715−3723.

(2) Evans, C. L.; Xie, X. S. Coherent Anti-Stokes Raman Scattering Microscopy: Chemical Imaging for Biology and Medicine. *Annu. Rev. Anal. Chem.* **2008**, *1*, 883−909.

(3) Min, W.; Freudiger, C. W.; Lu, S.; Xie, X. S. Coherent Nonlinear Optical Imaging: Beyond Fluorescence Microscopy. *Annu. Rev. Phys. Chem.* **2011**, *62*, 507−530.

(4) Garbacik, E.; Korterik, J.; Otto, C.; Mukamel, S.; Herek, J.; Offerhaus, H. L. Background-Free Nonlinear Microspectroscopy with Vibrational Molecular Interferometry. *Phys. Rev. Lett.* **2011**, *107*, 253902.

(5) Fussell, A.; Grasmeijer, F.; Frijlink, H.; de Boer, A.; Offerhaus, H. L. CARS microscopy as a tool for studying the distribution of micronised drugs in adhesive mixtures for inhalation. *J. Raman Spectrosc.* **2014**, *45*, 495−500.

(6) Cheng, J.-X.; Xie, X. S. Vibrational spectroscopic imaging of living systems: An emerging platform for biology and medicine. *Science* **2015**, *350*, aaa8870.

(7) Cleff, C.; Gasecka, A.; Ferrand, P.; Rigneault, H.; Brasselet, S.; Duboisset, J. Direct imaging of molecular symmetry by coherent anti-stokes Raman scattering. *Nat. Commun.* **2016**, *7*, 11562.

(8) Osseiran, S.; Wang, H.; Fang, V.; Pruessner, J.; Funk, L.; Evans, C. L. Nonlinear Optical Imaging of Melanin Species using Coherent Anti-Stokes Raman Scattering (CARS) and Sum-Frequency Absorp-

tion (SFA) Microscopy. *Optics in the Life Sciences Congress* **2017**, NS2C.3.
(9) Geissler, D.; Heiland, J. J.; Lotter, C.; Belder, D. Microchip HPLC separations monitored simultaneously by coherent anti-Stokes Raman scattering and fluorescence detection. *Microchim. Acta* **2017**, *184*, 315−321.
(10) Hirose, K.; Fukushima, S.; Furukawa, T.; Niioka, H.; Hashimoto, M. Invited Article: Label-free nerve imaging with a coherent anti-Stokes Raman scattering rigid endoscope using two optical fibers for laser delivery. *APL Photonics* **2018**, *3*, 092407.
(11) Karuna, A.; Masia, F.; Wiltshire, M.; Errington, R.; Borri, P.; Langbein, W. Label-Free Volumetric Quantitative Imaging of the Human Somatic Cell Division by Hyperspectral Coherent Anti-Stokes Raman Scattering. *Anal. Chem.* **2019**, *91*, 2813−2821.
(12) Levchenko, S. M.; Peng, X.; Liu, L.; Qu, J. The impact of cell fixation on coherent anti-stokes Raman scattering signal intensity in neuronal and glial cell lines. *Journal of Biophotonics* **2019**, *12*, e201800203.
(13) Nuriya, M.; Yoneyama, H.; Takahashi, K.; Leproux, P.; Couderc, V.; Yasui, M.; Kano, H. Characterization of Intra/Extracellular Water States Probed by Ultrabroadband Multiplex Coherent Anti-Stokes Raman Scattering (CARS) Spectroscopic Imaging. *J. Phys. Chem. A* **2019**, *123*, 3928−3934.
(14) Nishiyama, H.; Takamuku, S.; Oshikawa, K.; Lacher, S.; Iiyama, A.; Inukai, J. Chemical States of Water Molecules Distributed Inside a Proton Exchange Membrane of a Running Fuel Cell Studied by Operando Coherent Anti-Stokes Raman Scattering Spectroscopy. *J. Phys. Chem. C* **2020**, *124*, 9703.
(15) Rinia, H. A.; Bonn, M.; Müller, M. Quantitative Multiplex CARS Spectroscopy in Congested Spectral Regions. *J. Phys. Chem. B* **2006**, *110*, 4472−4479.
(16) Müller, M.; Rinia, H. A.; Bonn, M.; Vartiainen, E. M.; Lisker, M.; van Bel, A. Quantitative multiplex CARS spectroscopy in congested spectral regions. *Proc. SPIE* **2007**, 21−29.
(17) Rinia, H.; Burger, K. N. J.; Bonn, M.; Müller, M. Quantitative Label-Free Imaging of Lipid Composition and Packing of Individual Cellular Lipid Droplets Using Multiplex CARS Microscopy. *Biophys. J.* **2008**, *95*, 4908−4914.
(18) Day, J. P. R.; Domke, K. F.; Rago, G.; Kano, H.; Hamaguchi, H.-o.; Vartiainen, E. M.; Bonn, M. Quantitative Coherent Anti-Stokes Raman Scattering (CARS) Microscopy. *J. Phys. Chem. B* **2011**, *115*, 7713−7725.
(19) Vartiainen, E. M.; Rinia, H. A.; Müller, M.; Bonn, M. Direct extraction of Raman line-shapes from congested CARS spectra. *Opt. Express* **2006**, *14*, 3622−3630.
(20) Liu, Y.; Lee, Y. J.; Cicerone, M. T. Broadband CARS spectral phase retrieval using a time-domain Kramers−Kronig transform. *Opt. Lett.* **2009**, *34*, 1363−1365.
(21) Camp, C. H., Jr.; Lee, Y. J.; Cicerone, M. T. Quantitative, comparable coherent anti-Stokes Raman scattering (CARS) spectroscopy: correcting errors in phase retrieval. *J. Raman Spectrosc.* **2016**, *47*, 408−415.
(22) Kan, Y.; Lensu, L.; Hehl, G.; Volkmer, A.; Vartiainen, E. M. Wavelet prism decomposition analysis applied to CARS spectroscopy: a tool for accurate and quantitative extraction of resonant vibrational responses. *Opt. Express* **2016**, *24*, 11905−11916.
(23) Ababsa, F.; Mallem, M. Robust camera pose tracking for augmented reality using particle filtering framework. *Machine Vision and Applications* **2011**, *22*, 181−195.
(24) Liu, J.; Liu, D.; Dauwels, J.; Seah, H. S. 3D Human motion tracking by exemplar-based conditional particle filter. *Signal Processing* **2015**, *110*, 164−177.
(25) Moores, M. T.; Drovandi, C. C.; Mengersen, K.; Robert, C. P. Pre-processing for approximate Bayesian computation in image analysis. *Statist. Comput.* **2015**, *25*, 23−33.
(26) Lee, S.-H.; Kang, J.; Lee, S. Enhanced particle-filtering framework for vessel segmentation and tracking. *Computer Methods and Programs in Biomedicine* **2017**, *148*, 99−112.
(27) van Leeuwen, P. J. Particle Filtering in Geophysical Systems. *Mon. Weather Rev.* **2009**, *137*, 4089−4114.
(28) Effat, M. B.; Ciucci, F. Bayesian and Hierarchical Bayesian Based Regularization for Deconvolving the Distribution of Relaxation Times from Electrochemical Impedance Spectroscopy Data. *Electrochim. Acta* **2017**, *247*, 1117−1129.
(29) Edwards, T. H.; Stoll, S. A Bayesian approach to quantifying uncertainty from experimental noise in DEER spectroscopy. *J. Magn. Reson.* **2016**, *270*, 87−97.
(30) Yu, H.-F.; Dereli-Bégué, H.; Ryde, F. Bayesian Time-resolved Spectroscopy of GRB Pulses. *Astrophysical Journal* **2019**, *886*, 20.
(31) Bales, B.; Petzold, L.; Goodlet, B. R.; Lenthe, W. C.; Pollock, T. M. Bayesian inference of elastic properties with resonant ultrasound spectroscopy. *J. Acoust. Soc. Am.* **2018**, *143*, 71−83.
(32) Moores, M. T.; Gracie, K.; Carson, J.; Faulds, K.; Graham, D.; Girolami, M. *Bayesian modelling and quantification of Raman spectroscopy*. *arXiv* 2016; *1604.07299*.
(33) Kauppinen, J. K.; Moffatt, D. J.; Mantsch, H. H.; Cameron, D. G. Fourier Self-Deconvolution: A Method for Resolving Intrinsically Overlapped Bands. *Appl. Spectrosc.* **1981**, *35*, 271−276.
(34) Kauppinen, J. K.; Moffatt, D. J.; Hollberg, M. R.; Mantsch, H. H. A New Line-Narrowing Procedure Based on Fourier Self-Deconvolution, Maximum Entropy, and Linear Prediction. *Appl. Spectrosc.* **1991**, *45*, 411−416.
(35) Särkkä, S. *Bayesian Filtering and Smoothing*; Cambridge University Press, 2013.
(36) Chopin, N. A sequential particle filter method for static models. *Biometrika* **2002**, *89*, 539−552.
(37) Del Moral, P.; Doucet, A.; Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2006**, *68*, 411−436.
(38) Vartiainen, E. M. Phase retrieval approach for coherent anti-Stokes Raman scattering spectrum analysis. *J. Opt. Soc. Am. B* **1992**, *9*, 1209−1214.
(39) Cicerone, M. T.; Aamer, K. A.; Lee, Y. J.; Vartiainen, E. Maximum entropy and time-domain Kramers−Kronig phase retrieval approaches are functionally equivalent for CARS microspectroscopy. *J. Raman Spectrosc.* **2012**, *43*, 637−643.
(40) Mathlouthi, M.; Luu, D. V. Laser-raman spectra of d-fructose in aqueous solution. *Carbohydr. Res.* **1980**, *78*, 225−233.
(41) Douc, R.; Cappe, O. Comparison of resampling schemes for particle filtering. *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis* **2005**, 64−69.
(42) Söderholm, S.; Roos, Y. H.; Meinander, N.; Hotokka, M. Raman spectra of fructose and glucose in the amorphous and crystalline states. *J. Raman Spectrosc.* **1999**, *30*, 1009−1018.
(43) Brizuela, A. B.; Bichara, L. C.; Romano, E.; Yurquina, A.; Locatelli, S.; Brandán, S. A. A complete characterization of the vibrational spectra of sucrose. *Carbohydr. Res.* **2012**, *361*, 212−218.
(44) Kauppinen, J. K.; Moffatt, D. J.; Mantsch, H. H. Nonlinearity of the maximum entropy method in resolution enhancement. *Can. J. Chem.* **1992**, *70*, 2887−2894.
(45) Kauppinen, J. K.; Saario, E. K. What is Wrong with MEM? *Appl. Spectrosc.* **1993**, *47*, 1123−1127.

https://dx.doi.org/10.1021/acs.jpcb.0c04378
J. Phys. Chem. B 2020, 124, 7005−7012