Efficient quantitative hyperspectral image unmixing method for large-scale Raman micro-spectroscopy data analysis

E. G. Lobanova[a,*], S. V. Lobanov[b,c]

[a]*School of Biosciences, Cardiff University, Cardiff CF10 3AX, United Kingdom*
[b]*School of Physics and Astronomy, Cardiff University, Cardiff CF24 3AA, United Kingdom*
[c]*School of Medicine, Cardiff University, Cardiff CF24 4HQ, United Kingdom*

Abstract

Vibrational micro-spectroscopy is a powerful optical tool, providing a non-invasive label-free chemically specific imaging for many chemical and biomedical applications. However, hyperspectral image produced by Raman micro-spectroscopy typically consists of thousands discrete pixel points, each having individual Raman spectrum at thousand wavenumbers, and therefore requires appropriate image unmixing computational methods to retrieve non-negative spatial concentration and corresponding non-negative spectra of the image biochemical constituents. Here, we present a new efficient Quantitative Hyperspectral Image Unmixing (Q-HIU) method for large-scale Raman micro-spectroscopy data analysis. This method enables to simultaneously analyse multi-set Raman hyperspectral images in three steps: (i) Singular Value Decomposition with innovative Automatic Divisive Correlation which autonomously filters spatially and spectrally uncorrelated noise from data; (ii) a robust subtraction of fluorescent background from the data using a newly developed algorithm called Bottom Gaussian Fitting; (iii) an efficient Quantitative Unsupervised/Partially Supervised Non-negative Matrix Factorization method, which rigorously retrieves non-negative spatial concentration maps and spectral profiles of the samples' biochemical constituents with no *a priori* information *or* when one or several samples' constituents are known. As compared with state-of-the-art methods, our approach allows to achieve significantly more accurate results and efficient quantification with several orders of magnitude shorter computational time as verified on both artificial and real experimental data. We apply Q-HIU to the analysis of large-scale Raman hyperspectral images of human atherosclerotic aortic tissues and our results show a proof-of-principle for the proposed method to retrieve and quantify the biochemical composition of the tissues, consisting of both high and low concentrated compounds. Along with the established hallmarks of atherosclerosis including cholesterol/cholesterol ester, triglyceride and calcium hydroxyapatite crystals, our Q-HIU allowed to identify the significant accumulations of oxidatively modified lipids co-localizing with the atherosclerotic plaque lesions in the aortic tissues, possibly reflecting the persistent presence of inflammation and oxidative damage in these regions, which are in turn able to promote the disease pathology. For minor chemical components in the diseased tissues, our Q-HIU was able to detect the signatures of calcium hydroxyapatite and β-carotene with relative mean Raman concentrations as low as 0.09% and 0.04% from the original Raman intensity matrix with noise and fluorescent background contributions of 3% and 94%, respectively.

Keywords: Raman spectroscopy, Hyperspectral image analysis, Biochemical quantification, Baseline correction, Non-negative matrix factorization, Multivariate curve resolution

1. Introduction

Over the past few decades, the development in multivariate image analysis methods has paved the way to handle and interpret the biochemical information, received from spectroscopic images quite efficiently. For instance, spontaneous Raman micro-spectroscopy data have been treated using multivariate image reconstruction methods [1, 2], such as Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), and K-Means cluster Analysis (KMA) [3, 4, 5, 6, 7], utilizing spectral contrast originated from biochemical composition variations over sample image pixels in order to produce pseudo-color images revealing this contrast. However, in these methods the spatio-spectral information is sorted into components that do not represent individual biochemical species with physically meaningful spectra and concentration.

Raman data have also been analysed using Vertex Component Analysis (VCA) [8] and Multivariate Curve Resolution (MCR) [3, 9, 10, 11], which is also known as Non-negative Matrix Factorization (NMF) [12]. In contrast to VCA which requires the presence of pixels containing pure biochemical substances, MCR/NMF solves a general hyperspectral image unmixing problem without this restric-

*Corresponding author
Email addresses: LobanovaE@cardiff.ac.uk (E. G. Lobanova), LobanovS@cardiff.ac.uk (S. V. Lobanov)

tion and therefore has broader chemical and biomedical applications.

In relation to Coherent Raman Scattering (CRS) microspectroscopy [13, 14], Coherent Anti-Stokes Raman scattering (CARS) and Stimulated Raman scattering (SRS) images have been analysed by PCA [15], cluster analysis based on spectral phasor approach [16], MCR [17] and NMF [18] both based on the Alternating Non-negativity-constrained Least Squares algotithm (ANLS). In contrast to PCA and HCA, MCR/NMF can provide a quantitative determination of the chemical composition. Importantly, the popular MATLAB realization of the MCR-ANLS algorithm [19] uses computationally inefficient realization of the non-negative least squares (NLS) algorithm, which can be several orders of magnitude slower than fast combinatorial NLS (FC-NLS) [20, 21]. Indeed, the standard NLS algotithm [22] is designed to calculate a non-negative right hand-side (RHS) vector. The multiple-RHS problem, used to quantify the chemical information in a hyperspectral image, can be reduced to the independent solution of problems for each RHS vector. However, the computation of this technique is time consuming, which becomes even worse for big data-sets. For instance, the MCR-ANLS approach implemented by [19] has running times of days on a standard PC for a typical Raman data-set ($\simeq 50\times50\times10^3$ matrix) from a sample of unknown chemical composition, which requires in-depth analysis with the different number of components. For large volume data analysis with multi-set Raman images ($\simeq 10^3\times10^3\times10^3$ matrix) it leads to excessively long running times of years. To that end, there is a need for faster chemometric approaches, which reduce running times to hours for large data-sets without compromising the accuracy.

The goal of our paper is to develop efficient Quantitative Unsupervised/Partially Supervised Hyperspectral Image Unmixing (Q-HIU) method for the analysis of large-scale Raman micro-spectroscopy data.

2. Theory and algorithms

2.1. Formulation of hyperspectral image unmixing problem

Forward scattering problem. Suppose the examining specimens are a mixture of N pure chemicals (components) with known individual Raman scattering cross-section spectra[1] $S_i(\nu)$, $i = 1, 2, \ldots, N$ and known spatial distributions of concentration $C_i(\mathbf{r})$. Here, ν indicates Raman shift and $\mathbf{r} = (x, y; r)$ is a radius vector drawn to the image pixel with coordinates (x, y) of the r^{th} hyperspectral image. For this example problem, one could measure Raman intensity matrix $I(\mathbf{r}; \nu)$, which is a linear combination of the individual spatially-resolved concentrations and corresponding pure spectra products of the samples' constituents

$$I(\mathbf{r}; \nu) = \sum_{i=1}^{N} C_i(\mathbf{r}) S_i(\nu). \tag{1}$$

This equation represents the solution of the forward scattering problem, which is computationally straightforward.

Note, Eq. (1) can be also written in the matrix form as follows (see Supplementary material for details)

$$\mathbf{I} = \mathbf{C}^{\mathrm{T}}\mathbf{S}, \tag{2}$$

where the superscript $^{\mathrm{T}}$ means matrix transpose.

Inverse scattering problem. A more challenging problem, which is usually the case of chemical and biomedical applications, is to retrieve the chemical composition of samples without prior knowledge of their constituents. This inverse scattering problem can be formulated in a following way. For a given Raman intensity $I(\mathbf{r}; \nu)$, one should factorize it into N separate chemical components with individual Raman spectra $S_i(\nu)$ and spatially-resolved concentration profiles $C_i(\mathbf{r})$. In this paper, we also consider the case, where one or several Raman spectra $S_i(\nu)$ can be known.

It is important to note that since this problem retrieves unknown chemical composition from the Raman images, concentration of analytes can be only identified up to a constant factor, which can be determined in Raman experiment for each analyte individually. For this, one could measure Raman spectrum $S_i^0(\nu)$ of pure chemical compound (identified as a component from the Q-HIU analysis) with known chemical concentration c_i and multiply its concentration profile $C_i(\mathbf{r})$ by a calibration factor $c_i \int S_i(\nu)d\nu / \int S_i^0(\nu)d\nu$, which gives the chemical concentration of analyte. Since calibration factors are unknown for the Raman datasets to be investigated in this paper, we will use the term 'Raman concentration' instead of 'chemical concentration' to avoid confusion.

2.2. Singular Value Decomposition with Automatic Divisive Correlation (SVD-ADC) for noise filtering

A standard formulation of SVD algorithm [23] applied to hyperspectral intensity matrix $\mathbf{I}$ enables to project the data in a new spatio-spectral orthonormal basis consisting of scores $C_i(\mathbf{r})$ and loadings $S_i(\nu)$

$$I(\mathbf{r}; \nu) = \sum_{i\in\mathbb{N}} C_i(\mathbf{r}) \lambda_i S_i(\nu), \tag{3}$$

which can be sorted by their descending singular values λ_i [24]. Here, $\mathbb{N} = \{1, 2, \ldots, \mathrm{rank}(\mathbf{I})\}$ and $C_i(\mathbf{r}) \cdot C_j(\mathbf{r}) = S_i(\nu) \cdot S_j(\nu) = \delta_{ij}$, where the dots indicate Euclidean scalar products (see Supplementary material for definition).

It is believed that physically meaningful components have significantly higher singular values compared to noise ones. However, an adequate threshold criterion, which rejects relatively small singular values and its corresponding

[1] In the following, we will omit "scattering cross-section" for brevity.

singular vectors, to the best of our knowledge, does not exist in literature. Furthermore, singular vectors (scores and loadings) with small singular values can contain meaningful chemical information and, therefore, must not be discarded. Such meaningful components tend to have high values of autocorrelation function for small pixel shifts. For instance, if a chemical component with distinct Raman spectral profile (high autocorrelation) is localized in a few pixels, SVD might retrieve it as a singular component with small singular value and, therefore, it will potentially be filtered out. To this end, autocorrelation function appears to be more accurate for assessing singular components compared to singular values.

To address this problem, we propose SVD in a new formulation, which we call SVD with Automatic Divisive Correlation (SVD-ADC). This approach requires additional step once SVD is performed, i.e.: for each left and right singular vectors ($C_i(\mathbf{r})$ and $S_i(\nu)$, respectively) we calculate spectral R_i^S and spatial R_i^C autocorrelation coefficients at "one" pixel shift

$$R_i^S = R\left[S_i(\nu), S_i^{\delta}(\nu)\right], \tag{4}$$

$$R_i^C = \max_{\boldsymbol{\delta}} R\left[C_i(\mathbf{r}), C_i^{\boldsymbol{\delta}}(\mathbf{r})\right], \tag{5}$$

where $R[\ldots]$ denotes correlation coefficient (see Supplementary material for definition), $S_i^{\delta}(\nu)$ indicates loadings shifted by "one" spectral point, i.e. $S_i^{\delta}(\nu_j) = S_i(\nu_{j+\delta})$, and $C_i^{\boldsymbol{\delta}}(\mathbf{r})$ represents scores shifted by "one" image pixel along x- [$\boldsymbol{\delta} = (\delta_x, 0)$] or y-axis [$\boldsymbol{\delta} = (0, \delta_y)$]. Here, double quotation marks for the word "one" imply that one pixel shift should be chosen only when the spectral/spatial resolution of the used instrumentation is smaller than the pixel size of a CCD camera. For example, in our analysis presented below we used $\delta = \delta_x = \delta_y = 1$ for dataset 2 and $\delta = \delta_x = 3$, $\delta_y = 1$ for dataset 3.

This approach enables to automatically filter the meaningful components from noise-dominated ones (e.g.: shot noise, read noise, residual noise after cosmic rays removal, and etc.) by plotting the spatial autocorrelation coefficients R_i^C against spectral ones R_i^S for each pair of singular vectors and then discarding components with mean autocorrelation coefficient $R_i = (R_i^C + R_i^S)/2$ lower than $R_{\text{thr}} = 50\%$. This number was testified to adequately differentiate the meaningful components from the noise-dominated ones, which can be simply explained by the fact that an ideally noisy component exhibits 0% autocorrelation coefficient, whereas it is approximately 100% for an ideally meaningful component. After replacing the set $\mathbb{N}$ in Eq. (3) by the set $\mathbb{N}_{\text{thr}} = \{i\colon R_i > R_{\text{thr}}\}$ and performing summation over this set, we find Raman intensity matrix with reduced/removed noise.

2.3. Bottom Gaussian Fitting (BGF) for background subtraction

Raman spectra have an inherent broad background originating from fluorescence and amorphous scattering from glass substrate and sample itself, and therefore baseline correction algorithms via polynomial functions of different orders [25, 26] and Asymmetric Least Squares (AsLS) method [27, 28] are used for background removal. However, polynomial baseline correction methods might be hardly reproducible and give incorrect conclusions [29], whereas AsLS is believed to provide relatively accurate and automated background removal from hyperspectral Raman data [30]. Importantly, the AsLS algorithm requires the optimization over the smoothing λ and asymmetric p parameters, allowing broad background to be higher than total measured intensity which leads to negative non-physical Raman signal.

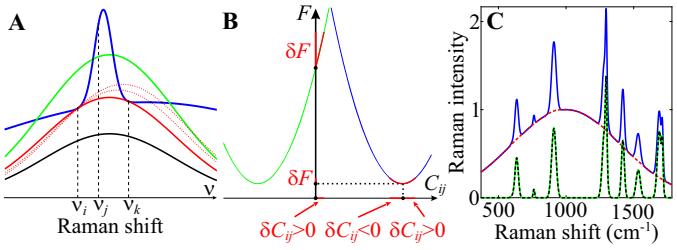

Figure 1: **A**, Schematic representation of the background subtraction procedure using BGF. The input artificial spectrum resembling Raman-like band is shown by the blue line. The black, red, and green lines show Gaussian functions with the same STDs σ, but different amplitudes A and expected values μ. **B**, Dependence of the factorization error $F(\mathbf{C}, \mathbf{S})$ on a matrix element C_{ij} for fixed other elements. Two special cases are shown: the blue parabola has a minimum at positive C_{ij} value, whereas a minimum of the green parabola is in the forbidden region $C_{ij} < 0$. Red lines indicate variation of the factorization error δF at the constrained minima: $C_{ij} > 0$ and $C_{ij} = 0$ for the first (blue parabola) and second (green parabola) cases, respectively. **C**, Example of the BGF procedure showing efficiency of the developed algorithm. The input artificial Raman-like spectrum $\mathbf{I}^{\text{a}} = \mathbf{I}^{\text{aR}} + \mathbf{I}^{\text{aB}}$ is shown by the blue line. The result of BGF $\mathbf{I}^{\text{R}}$ is shown by the black dashed line: the bottom Gaussian fit $\mathbf{I}^{\text{B}}$ (red dashed line) is subtracted from the Raman-like spectrum $\mathbf{I}^{\text{a}}$, resulting in the background-free Raman-like spectrum $\mathbf{I}^{\text{R}}$, which is in a good agreement with the model $\mathbf{I}^{\text{aR}}$ (green line).

To this end, we propose a new algorithm, Bottom Gaussian Fitting (BGF), allowing to autonomously subtract a complex curved background from Raman microspectroscopy data resulting in the accurate quantification of the Raman bands in the spectra at each image pixel. Compared to AsLS, our BGF benefits from only one intuitive parameter, which has units of Raman wavenumbers (cm^{-1}) and means minimum width of broad background features. The principles of this mathematical approach are explained in the following.

For the sake of brevity, we will omit in this section the first index in the Raman intensity matrix $\mathbf{I}$ labelling image pixel, i.e. we will write I_j instead of I_{ij} meaning that the background removing procedure must be done independently for each image pixel i.

Let us split the Raman intensity $\mathbf{I}$ into two non-negative parts

$$\mathbf{I} = \mathbf{I}^{\text{R}} + \mathbf{I}^{\text{B}}, \tag{6}$$

where $\mathbf{I}^{\text{R}}$ contains sharp resonance peaks representing Ra-

man bands, whereas $\mathbf{I}^{\mathrm{B}}$ is a quasi-slowly varied background approaching the Raman intensity $\mathbf{I}$ from the bottom. Our aim is to find $\mathbf{I}^{\mathrm{B}}$ as a quasi-superposition of Gaussian functions with standard deviations (STD) larger than σ – *the only one parameter* that will be used for background subtraction.

Let us consider some Raman shift ν_j and find maximum possible background intensity at this point I_j^{B}. The figure 1A illustrates a series of Gaussian functions (black, green, and red lines) with fixed STD σ and varying amplitude A and expectation value μ, which can potentially be background fits for a Raman spectrum $\mathbf{I}$ shown in Fig. 1A by the blue line. As one can see from the figure, only the red solid line tangents the Raman spectrum $\mathbf{I}$, so that the representative Raman intensity $\mathbf{I}^{\mathrm{R}}$ at the point ν_j is well quantified. Mathematically, since the background intensity $\mathbf{I}^{\mathrm{B}}$ tangents the Raman intensity $\mathbf{I}$, it should be determined as a Gaussian function $G_\sigma(\nu_j;\nu_i,I_i,\nu_k,I_k)$ with STD σ, which passes through left-sided (ν_i,I_i) and right-sided (ν_k,I_k) points of the Raman intensity $\mathbf{I}$ relating to the considered Raman shift ν_j. Furthermore, the tangent background fit $\mathbf{I}^{\mathrm{B}}$ (red solid line) has minimum Raman intensity value at the point ν_j as can be visually observed from the Fig. 1A (compare the red solid line with the red dotted and green lines). Thus, the bottom Gaussian fit can be defined as

$$I_j^{\mathrm{B}} = \min_{i\le j\le k} G_\sigma(\nu_j;\nu_i,I_i,\nu_k,I_k), \tag{7}$$

where the Gaussian function

$$G_\sigma(\nu_j;\nu_i,I_i,\nu_k,I_k) = Ae^{-\frac{(\nu_j-\mu)^2}{2\sigma^2}} \tag{8}$$

has amplitude A and expected value μ, which are implicitly defined by the following equations

$$Ae^{-\frac{(\nu_i-\mu)^2}{2\sigma^2}} = I_i, \quad Ae^{-\frac{(\nu_k-\mu)^2}{2\sigma^2}} = I_k. \tag{9}$$

Computing Eq. (7) for all Raman shifts ν_j, we find a quasi-slowly varying curve $\mathbf{I}^{\mathrm{B}}$ fitting the Raman intensity $\mathbf{I}$ from the bottom. The difference between these two curves ($\mathbf{I}^{\mathrm{R}} = \mathbf{I} - \mathbf{I}^{\mathrm{B}}$) contains only sharp resonances or Raman bands.

Once the background-free Raman-like signal $\mathbf{I}^{\mathrm{R}}$ is retrieved using BGF, we calculate the spatially-resolved spectral auto-correlation coefficients for this matrix at "one" pixel shift and then discard the spatial points, which have correlation values lower than $R^S_{\mathrm{thr}} = 50\%$, from the further analysis. This optional additional step allows to eliminate contribution of spatial points, which are strongly affected by noise due to dominated background (orders of magnitude larger than the remaining signal).

Note that Eq. (7) requires two loops (over the indexes i and k), which is apparently computationally time consuming. However, it can be simply reduced to one loop, which is utilized in our BGF method.

2.4. Efficient Quantitative Unsupervised/Partially Supervised Non-negative Matrix Factorization (Q-US/PS-NMF)

In a standard formulation, the NMF approach [31, 32, 33, 34] allows to decompose the spatially-resolved Raman intensity matrix $\mathbf{I}$ into a product Eq. (2) of non-negative spatial concentration maps $\mathbf{C}$ and corresponding non-negative spectra $\mathbf{S}$ with rows representing individual biochemical substances (components) of the samples' composition. Both matrices $\mathbf{C}$ and $\mathbf{S}$ are unknown prior to NMF.

In this paper, we generalise NMF to operate as a partially supervised method, which we call Q-US/PS-NMF. This modality can be particularly useful when required to eliminate the contribution of wax residues to the spectral basis for paraffin-embedded samples. For this, the spectrum of paraffin-wax compound is fixed during the Q-US/PS-NMF procedure.

Mathematically, the problem can be formulated as follows. Suppose N_k biochemical substances with Raman spectra $S_i(\nu)$, $i = 1,\ldots,N_k$ are known and our goal is to determine the rest $N_u = N - N_k$ spectra $S_i(\nu)$, $i = N_k+1,\ldots,N$, as well as concentration maps for both known and unknown components $C_i(\mathbf{r})$, $i = 1,2,\ldots,N$. Denoting known/unknown parts of the spectral matrix $\mathbf{S}$ and corresponding parts of the concentration matrix $\mathbf{C}$ with letters k/u, we have

$$\mathbf{C} = \begin{pmatrix}\mathbf{C}_k\\ \mathbf{C}_u\end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix}\mathbf{S}_k\\ \mathbf{S}_u\end{pmatrix}. \tag{10}$$

Ideally, these matrices should satisfy Eq. (2). However, for real experimental data, the Raman intensity matrix $\mathbf{I}$ can be factorized only approximately with a residue $\mathbf{E}$, i.e. $\mathbf{I} - \mathbf{C}^{\mathrm{T}}\mathbf{S} = \mathbf{E}$. A pair of matrices $\mathbf{C}$ and $\mathbf{S}$, which minimises Frobenius norm of the residue $\|\mathbf{E}\|_{\mathrm{F}}$, represents the solution of the factorization problem Eq. (2). This is equivalent to the minimization of the following function

$$F(\mathbf{C},\mathbf{S}) = \frac{1}{2}\|\mathbf{I} - \mathbf{C}^{\mathrm{T}}\mathbf{S}\|^2_{\mathrm{F}} \tag{11}$$

subject to $C_{ij} \ge 0$, $S_{ij} \ge 0$. The necessary condition for a minimum of Eq. (11) is that a variation of the factorization error δF at the point $(\mathbf{C},\mathbf{S})$ is non-negative, i.e.

$$\delta F = \sum_{i=1}^{N}\sum_{j=1}^{N_p}\frac{\partial F}{\partial C_{ij}}\delta C_{ij} + \sum_{i=N_k+1}^{N}\sum_{j=1}^{N_s}\frac{\partial F}{\partial S_{ij}}\delta S_{ij} \ge 0. \tag{12}$$

Note, since variations δC_{ij} and δS_{ij} are arbitrary, each term of this equation must be greater or equal to zero.

Let us consider dependence of the factorization error F on a variable C_{ij} for fixed other variables. The graph of this function is parabola, which can have the minimum either at a point $C_{ij} > 0$ (blue line in Fig. 1B) or in the forbidden region $C_{ij} < 0$ (green line in Fig. 1B). In the former, variation δC_{ij} (highlighted by red color in Fig. 1B) can be both positive and negative, which requires

the partial derivative $\partial F/\partial C_{ij}$ to be zero in order to satisfy Eq. (12). In the latter, the constrained minimum must be shifted from the region with negative values of C_{ij} to the nearest possible point in the non-negative region, which is $C_{ij} = 0$, so the variation δC_{ij} can have only positive values. Thus, the partial derivative $\partial F/\partial C_{ij}$ is not required to be zero and can take positive values. These requirements are called the Karush-Kuhn-Tucker conditions [22] and can be written in a following way

$$\frac{\partial F}{\partial C_{ij}} = 0 \text{ if } C_{ij} > 0, \qquad \frac{\partial F}{\partial C_{ij}} \geq 0 \text{ if } C_{ij} = 0, \quad (13)$$

$$\frac{\partial F}{\partial S_{ij}} = 0 \text{ if } S_{ij} > 0, \qquad \frac{\partial F}{\partial S_{ij}} \geq 0 \text{ if } S_{ij} = 0, \quad (14)$$

where the bottom formulas must be valid only for $i > N_k$, i.e. for unknown spectra.

Differentiating Eq. (11), we find partial derivatives

$$\frac{\partial F}{\partial \mathbf{C}} = \mathbf{S}\mathbf{S}^{\mathrm{T}}\mathbf{C} - \mathbf{S}\mathbf{I}^{\mathrm{T}}, \quad (15)$$

$$\frac{\partial F}{\partial \mathbf{S}_u} = \mathbf{C}_u\mathbf{C}_u^{\mathrm{T}}\mathbf{S}_u - \mathbf{C}_u\mathbf{I} + \mathbf{C}_u\mathbf{C}_k^{\mathrm{T}}\mathbf{S}_k, \quad (16)$$

which lead to a non-linear problem after substitution them into Eqs. (13)-(14). This non-linear problem can be solved iteratively by alternative fixation of one matrix ($\mathbf{C}$ or $\mathbf{S}_u$) and computation of another one ($\mathbf{S}_u$ or $\mathbf{C}$, respectively) using NLS algorithm [22]. Namely, alternately denoting $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ and $\mathbf{C}_u\mathbf{C}_u^{\mathrm{T}}$ by $\mathbf{A}$, $\mathbf{S}\mathbf{I}^{\mathrm{T}}$ and $\mathbf{C}_u\mathbf{I} - \mathbf{C}_u\mathbf{C}_k^{\mathrm{T}}\mathbf{S}_k$ by $\mathbf{B}$, unknown non-negative matrices $\mathbf{C}$ and $\mathbf{S}_u$ by $\mathbf{X}$, and independently considering each column $\mathbf{x}$ and $\mathbf{b}$ of the matrices $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots)$ and $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots)$, one find a solution of the NLS problem

$$\begin{matrix} \mathbf{x}_p = \mathbf{A}_{pp}^{-1}\mathbf{b}_p \\ \mathbf{x}_a = 0 \end{matrix} \quad \text{subject to} \quad \begin{matrix} \mathbf{x}_p > 0, \\ \mathbf{A}_{ap}\mathbf{x}_p - \mathbf{b}_a \geq 0. \end{matrix} \quad (17)$$

Here, the indexes p and a denote projection of the vectors $\mathbf{x}$, $\mathbf{b}$ and matrix $\mathbf{A}$ onto passive and active subspaces characterizing by two complementary sets $\mathbb{P}$ and $\mathbb{A}$, respectively:

$$\begin{matrix} \mathbb{P} \cap \mathbb{A} = \varnothing, & \mathbb{P} \cup \mathbb{A} = \{1, 2, \ldots\}, \\ \mathbf{x}_p = \{x_i \colon i \in \mathbb{P}\}, & \mathbf{x}_a = \{x_i \colon i \in \mathbb{A}\}, \\ \mathbf{A}_{pp} = \{A_{ij} \colon i \in \mathbb{P},\, j \in \mathbb{P}\}, & \mathbf{A}_{ap} = \{A_{ij} \colon i \in \mathbb{A},\, j \in \mathbb{P}\} \end{matrix}$$

and similar for $\mathbf{b}$.

The described procedure can be done independently for each column of the matrix $\mathbf{X}$. However, working with the whole matrices rather than vectors is more computationally efficient [20, 21]. Indeed, one can group the columns that share the same index sets $\mathbb{P}$ and compute the appropriate matrix inverse $\mathbf{A}_{pp}^{-1}$ in Eq. (17) for each block. This procedure is apparently orders of magnitude faster than sequential computation of the matrix inverse for each individual column (see Secs. 4.2.3 and 4.3.4 for detail), which is used in the popular MATLAB realization of the MCR-ANLS algorithm by [19] (see Table 1 and Fig. S1 of the Supplementary material).

Note that Eqs. (15)-(16) can be generalized to the case of partially known spatial distributions of concentration, which might be relevant for other scientific applications. To this end, the term $\mathbf{S}_u\mathbf{S}_k^{\mathrm{T}}\mathbf{C}_k$ should be added to the right part of Eq. (15).

Furthermore, the Q-US/PS-NMF algorithm allows to quantitatively determine and compare pixel-by-pixel Raman concentrations of different biochemical components between samples by applying the following normalization conditions during final step of Q-US/PS-NMF analysis: the spectrum of each factorized component is normalized on *const* so that the mean spectral intensity for each component was the same and the sum of components' mean Raman concentrations is set equal to one.

$$\int d\nu S_1(\nu) = \int d\nu S_2(\nu) = \ldots = \int d\nu S_N(\nu), \quad (18)$$

$$\frac{1}{N_p}\sum_{j=1}^{N_p}\sum_{i=1}^{N} C_i(\mathbf{r}_j) = 1, \quad (19)$$

where N_p is a total number of pixel points.

Overall, Q-HIU presented here enables to decompose large Raman multi-datasets, simultaneously analysed from a variety of samples, into individual chemical components with spatially-resolved concentration and spectral profiles, in a quantitative manner without prior knowledge of the chemical composition of samples or with partial knowledge of some constituents' spectra.

3. Materials and methods

3.1. Software

Our Q-HIU software used to produce the results of this paper and containing three newly-developed functions: SVD-ADC, BGF, and Q-US/PS-NMF intended for the quantitative analysis of Raman microscopy images, is freely available from https://github.com/LobanovaEG-LobanovSV/Q-HIU.git.

3.2. Datasets

Dataset 1 consists of a superposition of Gaussian functions resembling Raman-like signal $\mathbf{I}^{\mathrm{aR}}$ superimposed by a background $\mathbf{I}^{\mathrm{aB}}$. The background has unitary amplitude $A^{\mathrm{aB}} = 1$, STD $\sigma^{\mathrm{aB}} = 400\,\mathrm{cm}^{-1}$, and expected value $\mu^{\mathrm{aB}} = 1000\,\mathrm{cm}^{-1}$. The amplitudes, STDs, and expected values for the ten Raman-like bands were randomly selected from the intervals $(0, 1)$, $(5, 20)\,\mathrm{cm}^{-1}$, and $(370, 1783)\,\mathrm{cm}^{-1}$, respectively.

Dataset 2 consists of mixtures of Raman spectra of 8 pure chemical compounds (sphingomyelin, β-carotene, collagen, cholesteryl palmitate, elastin, iron transferrin, cholesterol, and iron (III) oxide) with simulated random spatial distribution of concentration (see Supplementary Figs. S31-S38). The dataset was superimposed by a white Gaussian noise and a background. The amplitude of the white Gaussian noise was chosen to have a dynamic range

of signal-to-noise ratio SNR $= (0.01, 10^{10})$. The background was modelled as a Gaussian function, which has a spectrum and simulated spatial distribution of concentration as shown in Supplementary Fig. S30.

Multi-dataset 3 consists of 3 large volume Raman images (see schematic of a hyperspectral data cube in Fig. 4A) of diseased aortic tissues, containing biochemical signatures of atherosclerotic plaque lesions (e.g.: foam cells/necrotic core, cholesterol crystals, calcification), and 3 Raman images of non-atherosclerotic age-matched control ones. In the following, we refer to atherosclerotic (A) and non-atherosclerotic control (C) samples as A1 (022_p_000), A2 (025_p_000), A3 (031_p_005), and C1 (043_000), C2 (046_000), C3 (047_002), respectively (see [35] for details on tissue cohorts).

4. Results and discussion

In Sec. 2, we presented a novel efficient Q-HIU approach for large-scale Raman micro-spectroscopy data analysis, consisting of three methods: SVD-ADC, BGF, Q-US/PS-NMF, each of which represents main step in analysis. In the following, we demonstrate three examples of the Q-HIU analysis applied to both simulated and real experimental Raman micro-spectroscopy data.

4.1. Validation of BGF on artificial data (Dataset 1)

To show the efficiency of the developed algorithm, we give an example of the BGF procedure applied to an artificial spectrum displaying ten Raman bands (Fig. 1C). This input spectrum was chosen to be a superposition of Gaussian functions resembling Raman-like signal $\mathbf{I}^{\mathrm{aR}}$ superimposed by a background $\mathbf{I}^{\mathrm{aB}}$ with parameters as detailed in Sec. 3.2. As one can see from Fig. 1C, the background-free Raman-like signal $\mathbf{I}^{\mathrm{R}}$ (black dashed line), resulted from background subtraction with the representative bottom Gaussian fit $\mathbf{I}^{\mathrm{B}}$ of $\sigma = 300\,\mathrm{cm}^{-1}$ (red dashed line), is in a good agreement with the original Raman-like signal $\mathbf{I}^{\mathrm{aR}}$ (green line).

4.2. Validation of Q-HIU on mixtures of pure biochemical references (Dataset 2)

4.2.1. Noise filtering: SVD-ADC

To show the efficiency of the SVD-ADC algorithm, we add white Gaussian noise $\mathbf{I}^{\mathrm{n}}$ to the intensity matrix of biochemical mixtures of pure Raman spectra $\mathbf{I}^{\mathrm{mR}}$ and apply our noise filtering procedure *with no input parameters* to the resulting matrix $\mathbf{I} = \mathbf{I}^{\mathrm{mR}} + \mathbf{I}^{\mathrm{n}}$. Fig. 2A shows dependence of noise removal factor (NRF) on SNR, where

$$\mathrm{SNR} = \frac{\|\mathbf{I}^{\mathrm{mR}}\|_{\mathrm{F}}^2}{\|\mathbf{I}^{\mathrm{n}}\|_{\mathrm{F}}^2} \tag{20}$$

and NRF is defined as a relative difference between the denoised $\mathbf{I}^{\mathrm{R}}$ and reference $\mathbf{I}^{\mathrm{mR}}$ matrices normalized on noise power

$$\mathrm{NRF} = \frac{\|\mathbf{I}^{\mathrm{R}} - \mathbf{I}^{\mathrm{mR}}\|_{\mathrm{F}}^2}{\|\mathbf{I}^{\mathrm{n}}\|_{\mathrm{F}}^2}. \tag{21}$$

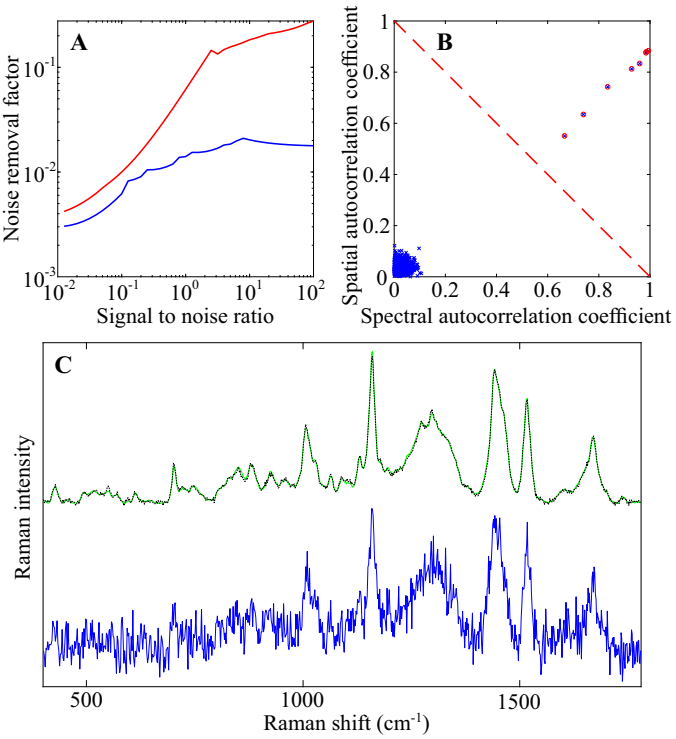

Figure 2: Noise removal efficiency of SVD-ADC applied to biochemical mixtures of pure Raman spectra with white Gaussian noise (dataset 2). **A**, Dependence of noise removal factor (NRF) on signal-to-noise ratio (SNR) for SVD-ADC (blue line) and MNF combined with S-G filtering (red line). **B**, Spatio-spectral autocorrelation coefficients map of singular vectors found from SVD-ADC for SNR = 10. The dashed diagonal line represents a decision line for mean autocorrelation coefficients R_i at $R_{\mathrm{thr}} = 50\%$, separating the coefficients above the line (circled cross signs – meaningful components) from those below it (cross signs – noise). **C**, Comparison of Raman spectrum *after* SVD-ADC procedure ($\mathbf{I}^{\mathrm{R}}$, black dashed line) with the known reference ($\mathbf{I}^{\mathrm{mR}}$, green line) for SNR = 10. The input Raman spectrum superimposed by white Gaussian noise is shown below ($\mathbf{I}$, blue line).

As can be seen from Fig. 2A, SVD-ADC allows to reduce noise from Raman signal by two orders of magnitude (NRF $\simeq 0.01$).

In order to illustrate the performance of SVD-ADC, we show the distribution of spatio-spectral autocorrelation coefficients for singular vectors found from SVD-ADC for SNR = 10 on Fig. 2B. Visual inspection verifies that the singular vectors representing noise are localised around the origin of coordinates, whereas the singular vectors attributed to meaningful components (circled cross signs) are in the upper-right half plane. Therefore, the dashed diagonal line with $R_{\text{thr}} = 50\%$ separates the meaningful components from noise-dominated ones. Fig. 2C also confirms that the denoised Raman spectrum $\mathbf{I}^{\text{R}}$ (black dashed line) is almost identical to the reference one $\mathbf{I}^{\text{mR}}$ (green line), when the noise in the input Raman spectrum (blue line) is comparable with the reference Raman signal.

In order to show the accuracy of the proposed method, we compare our SVD-ADC with the Maximum Noise Fraction (MNF) method [36] combined with Savitzky-Golay (S-G) filtering [37]. For this, we applied MNF on the same hyperspectral dataset of known chemical mixtures superimposed by white Gaussian noise and retained N eigenimages sorted according to decreasing SNR or reducing image quality. The rest eigenimages were denoised using S-G filtering with two smoothing parameters: polynomial order and frame length. In brief, we performed optimization over these 3 parameters for each SNR, which minimizes denoising error. As the final step, the calculation of inverse MNF on reduced noise eigenimages allowed to produce the resulting denoised data. Fig. 2A shows that both methods give similar NRF for noise-dominated data with SNR from the range $(0.01, 1)$. However, when SNR is higher than 10, MNF reduces noise by only 80% (NRF = 20%), whereas SVD-ADC allows to achieve 98% of noise reduction (NRF = 2%). Moreover, the determination of optimal parameters for MNF and S-G filtering is possible *only* when the true solution is known. In contrast, our SVD-ADC method allows to autonomously reduce noise from the hyperspectral Raman data with *no input* parameters.

4.2.2. Background subtraction: BGF

To validate the accuracy of BGF, we compare it with AsLS algorithm [27, 28] using biochemical mixtures of pure Raman spectra $\mathbf{I}^{\text{mR}}$ with Gaussian background. We investigated the dependence of background removal relative error

$$E = \frac{\max\limits_{ij} |I^{\text{R}}_{ij} - I^{\text{mR}}_{ij}|}{\max\limits_{ij} I^{\text{mR}}_{ij}} \qquad (22)$$

on corresponding parameters of BGF (σ, blue line with crosses) and AsLS (λ and fixed $p = 10^{-3}$, red line with pluses) and the results reveal that BGF allows to achieve relative error of 7%, whereas AsLS gives relative error of 64% (see Fig. 3A). Here, $\mathbf{I}^{\text{R}}$ is a background-free Raman intensity matrix retrieved from BGF or AsLS. Examples of Raman spectra (blue lines) at two spatial pixels

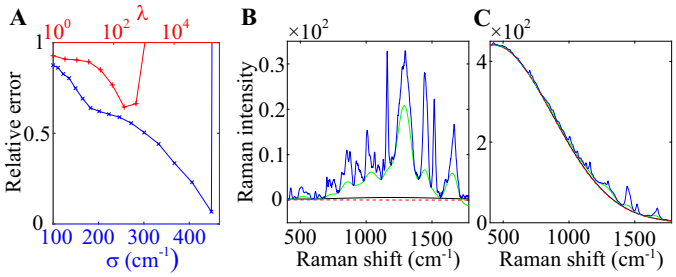

Figure 3: Comparison of BGF with AsLS on biochemical mixtures of pure Raman spectra with Gaussian background. **A**, Dependence of background removal relative error on corresponding parameters of BGF (σ, blue line with crosses) and AsLS (λ, red line with pluses). **B-C**, The Raman spectra (blue lines) at two spatial pixels with low (B) and high (C) true background (red dashed lines). The bottom Gaussian fits defined from BGF and baselines found from AsLS are shown by black and green lines, respectively. The panels B-C are shown for optimal parameters found from the panel A.

Table 1: Comparison of relative factorization error of Q-US/PS-NMF, MCR-ALS, and VCA for the 32×32×804 data matrix (dataset 2) performed on Intel Core i5-2537M

CPU time (min)	Q-US/PS-NMF	MCR-ALS	VCA
0	70%		70% (0.4 sec)
1	2.6%	46%	
5	0.7%	17%	
10	0.5%	11%	

with low (B) and high (C) true background (red dashed lines) show that BGF works equally well for low/high background (black lines), whereas AsLS may produce the baseline (green lines), which significantly overfits the original Raman signal and also has non-physical negative values (see Fig. 3B-C). Note that Fig. 3B-C are shown for optimal parameters found from Fig. 3A. These are $\sigma = 449\,\text{cm}^{-1}$ for BGF and $\lambda = 223, p = 10^{-3}$ for AsLS.

4.2.3. Q-US/PS-NMF unmixing procedure

To validate the accuracy of Q-US/PS-NMF, we compare it with the MCR-ALS algorithm [19] and VCA [38] using biochemical mixtures of 8 pure Raman spectra (32×32×804 data matrix). The results reveal that VCA is unable to retrieve the original chemical components and produces the relative factorization error of 70%. Also, our Q-US/PS-NMF shows fast convergence to the true solution and allows to achieve relative decomposition error of 2.6%, 0.7%, and 0.5% for 1, 5, and 10 minutes of standard PC running times, respectively, which are about 20 times accurate compared to MCR-ALS (see Table 1).

4.3. Validation of Q-HIU on real data (Dataset 3)

In this section, we apply the Q-HIU method to the large-scale analysis of real experimental Raman microspectroscopy images of human thoracic aortic tissues, which were sourced from [39] (see description of the data in Sec. 3.2), and compare its performance with existing

state-of-the-art methods. The Q-HIU analysis was performed *simultaneously* on all Raman images, enabling to show the computational efficiency of the developed approach. The resulting hyperspectral Raman multi-image was a $\simeq 6\times10^2\times10^3\times10^3$ matrix, where the first, second-third, and fourth dimensions represent the numbers of participating samples, spatial pixels in x- and y-directions, and spectral points, respectively.

Importantly, the new developed Q-US/PS-NMF method in a partially supervised formulation also presented in this paper is already used by us and co-authors in the quantitative analysis of large-scale label-free Raman hyperspectral images from biomedical paraffin-embedded samples in Alzheimer's disease [40]. In that paper, wax component was fixed during the Q-US/PS-NMF procedure in order to avoid the spectral contamination of the biochemical components, characteristic for human brain tissue, by wax residues.

4.3.1. Noise filtering: SVD-ADC

As well-documented in the literature, conventional SVD procedure for spectral image processing allows to separate signal from noise by rejecting small singular values and their corresponding singular vectors from the further analysis. The SVD-ADC approach, proposed in this paper, utilises autocorrelation coefficients of spatial and spectral singular vectors for noise filtering and allows to autonomously remove noise from Raman data with no input parameters. In particular, for each pair of singular vectors we calculate spatial and spectral autocorrelation coefficients and discard the singular vectors, which have a mean autocorrelation coefficient smaller than 50%.

Fig. 4D shows the autocorrelation coefficients map for singular vectors received from the SVD-ADC analysis of 6 Raman images. The dashed diagonal line represents a decision line at $R_{\text{thr}} = 50\%$ cut-off, that separates the coefficients above the line (meaningful components) from those below it (noise). Note that read-noise and shot noise are localized around zero, whereas physically meaningful components show high mean autocorrelation.

Importantly, the new approximating matrix, received after SVD on given data, is compressed (or has *significantly* reduced-rank), therefore it is important to carefully choose the cut-off value and avoid rejection of meaningful components. For our data example, SVD-ADC with a 50% cut-off of mean autocorrelation coefficients has found the approximating matrix, consisting of 19 singular components. To verify this number, we show the spatial distributions and corresponding spectra for two singular vectors with mean autocorrelation coefficients, which are 20% lower and higher than 50% value (Fig. 4E,F). For a pair of *discarded* singular vectors, corresponding to 30% mean autocorrelation, spatial maps do not show any recognised pattern and spectrum resembles noise, whereas for a pair of meaningful singular vectors with 70% mean autocorrelation, spatial maps show some distinct features and spectrum exhibits several clear Raman bands.

4.3.2. Background subtraction: BGF

Fig. 4B,C shows an example of our BGF background subtraction algorithm applied to the Raman data, found from SVD-ADC as detailed in the previous section. The Raman spectra at two spatial pixels (high and low background) from the selected hyperspectral image *before* (black line) and *after* (green line) background removal with the representative bottom Gaussian fit (red dashed line) are shown.

4.3.3. Q-US/PS-NMF unmixing procedure and interpretation of results

After SVD-ADC filtering and BGF background subtraction, we analysed the Raman data using Q-US/PS-NMF in the fingerprint ($370-1783\,\text{cm}^{-1}$) region. The Q-US/PS-NMF analysis found that the given 6 Raman images were optimally described using 11 separate components ($\mathcal{C}$). To validate this component number selection, we performed Q-US/PS-NMF with the numbers of components varied from 1 to 18, and then investigated the evolution of resulting component spectra (see Fig. 5D), which includes the formation of new biochemical components (red numbers in circles) and their splitting/mixing. The degree of component similarity between two analyses with subsequent numbers of components is indicated by solid (strong), dashed (medium), and dotted (weak) lines. For example, the formation of $\mathcal{C}_1$ and $\mathcal{C}_{10}$ in 11-component analysis, which will be later attributed to actin and hydroxyapatite, respectively, was a result of splitting of $\mathcal{C}_9$ in 10-component analysis. Note the components are sorted by their mean Raman concentration, from largest to smallest. The dependence of the relative factorization error on the number of components was also inspected. The results (see Fig. S5 of the Supplementary material) show a gradual decrease of this error when increasing the number of components, and therefore imply that the factorization error can not be used to determine the optimal number of components in the Q-US/PS-NMF data analysis.

7 components were found to be spectrally assigned to actin ($\mathcal{C}_1$, $R^2 = 85\%$); elastin ($\mathcal{C}_3$, $R^2 = 84\%$); a mixture of phospholipids and triglycerides ($\mathcal{C}_4$, $R^2 = 93\%$) at a ratio of about 2:1 (Fig. 5A); a mixture of cholesterol, cholesteryl esters with saturated fatty acid chains, and oxidized all-*trans* retinol also known as vitamin A ($\mathcal{C}_5$, $R^2 = 94\%$) at ratios of about 6:5:1 (Fig. 5B); collagen ($\mathcal{C}_6$, $R^2 = 83\%$); hydroxyapatite ($\mathcal{C}_{10}$, $R^2 = 93\%$); a mixture of β-carotene and cholesterol ($\mathcal{C}_{11}$, $R^2 = 88\%$) at a ratio of about 3:1 (see the component spectra in Figs. S6-S16 of the Supplementary material). Here, the contributions of biochemical reference components into each component spectra was determined using NLS fitting algorithm. The degree of similarity between each component spectra and its biochemical model was indicated by a correlation coefficient (R^2). The model uses the Raman spectra of the chemical species of analytical standards [35, 41, 42]. Additionally, $\mathcal{C}_2$ was found to be consistent with the Raman spectrum of water ($R^2 = 94\%$) using an ID expert tool of Bio-Rad's

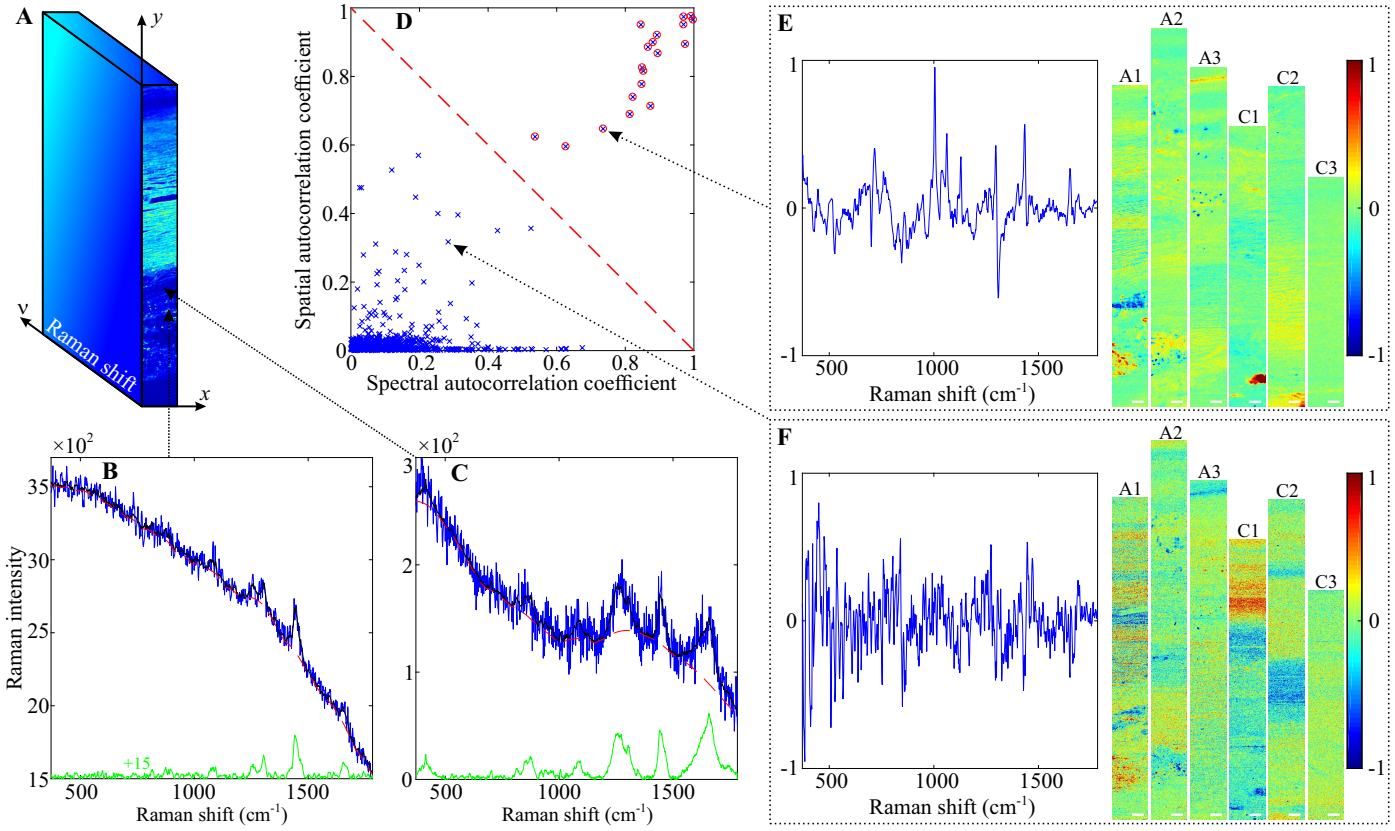

Figure 4: **A**, Schematic of a hyperspectral data cube. The arrows mark two pixels, Raman spectra of which are shown on the panels **B** and **C**. **B-C**, The input Raman spectrum (blue line) *before* SVD-ADC and the corresponding noise-filtered Raman spectrum (black line) *after* this procedure. The results of background subtraction using BGF is shown by the green line: the bottom Gaussian fit (red dashed line) is subtracted from the noise-filtered Raman spectrum, resulting in background-free Raman spectrum. **D**, As Fig. 2B. **E-F**, The spectra and spatial distributions of two singular vectors with the mean autocorrelation coefficients of (50±20)% as indicated on the panel **D**. On both panels, 6 Raman images are labelled according to the sample source. Scale bars are 100 μm.

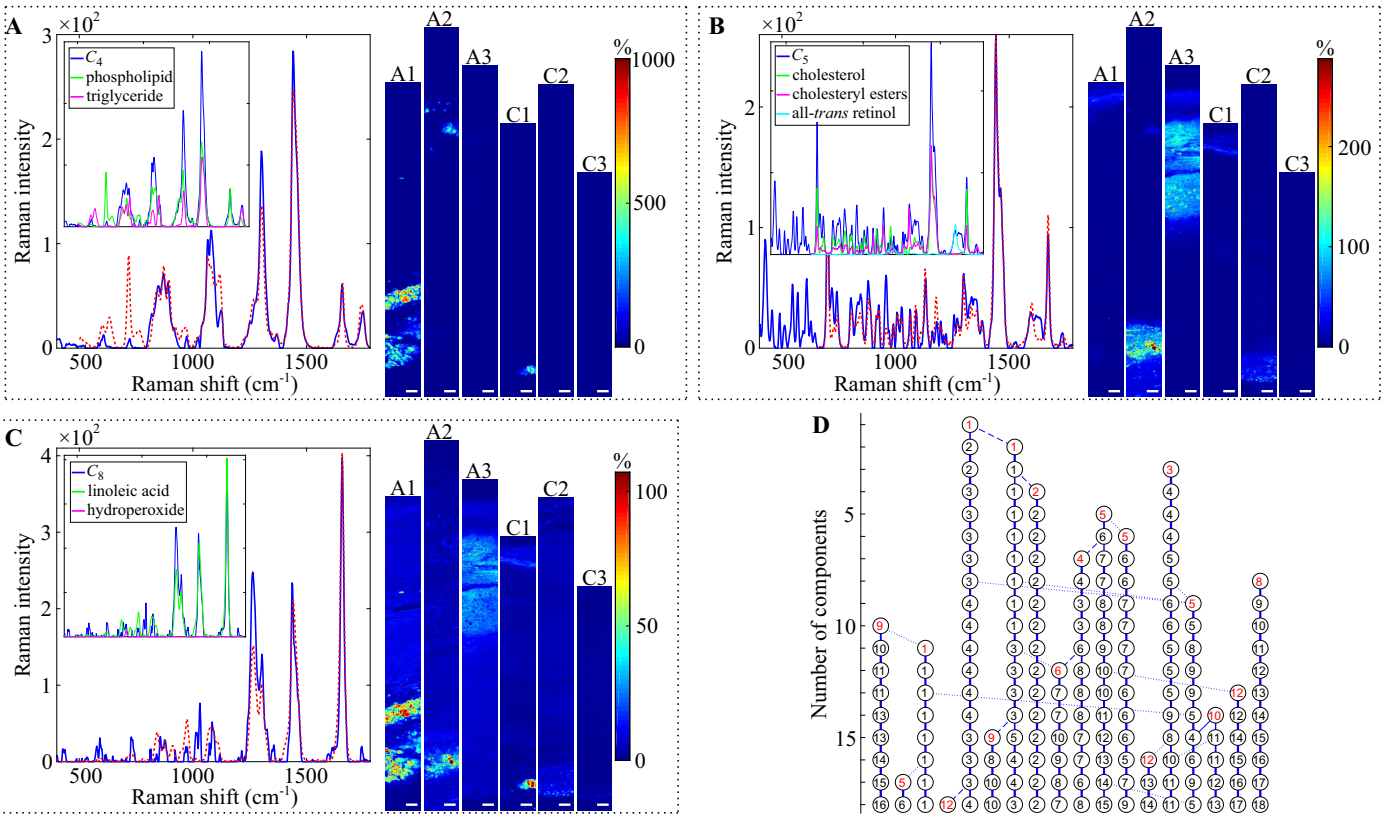

Figure 5: **A**-**C**, Component spectra (blue lines) and spatial distributions of Raman concentration of $\mathcal{C}_4$, $\mathcal{C}_5$, and $\mathcal{C}_8$ found from Q-US/PS-NMF of 6 Raman images from atherosclerotic aortic tissue regions (A1, A2, A3) and non-atherosclerotic controls (C1, C2, C3). The chemical attribution of components is based on the comparison with analytical standard Raman spectra of pure chemical species. Fits are shown by red dashed lines. Partial contributions (green, magenta, and cyan lines) to each component spectra (blue lines) are also indicated in the inset. Scale bars are 100 μm. **D**, Evolution of component spectra resulting from Q-US/PS-NMF with the numbers of components varied from 1 to 18, showing the appearance of new biochemical components (red numbers in circles) and their splitting/mixing. The degree of component similarity between two analysis with subsequent numbers of components is indicated by solid (strong), dashed (medium), and dotted (weak) lines.

KnowItAll Vibrational Spectroscopy software with Raman Spectral Libraries.

Generally, the results of the Q-US/PS-NMF analysis are consistent with the published results found from the VCA analysis [35]. However, our method allowed to retrieve additional biochemical information on the aortic tissue composition. In particular, it reveals a new biochemical component, which is spectrally attributed to oxidized linoleic acid ($\mathcal{C}_8$, $R^2 = 90\%$) (Fig. 5C). Importantly, the spectrum of this component (blue line) shows two unique Raman bands at 870 cm^{-1} and 1744 cm^{-1}, assigned to the -O-OH stretching mode, characteristic of hydrogen peroxide (magenta line in inset), and C=O stretching vibrations, which both occur in the process of lipid peroxidation. Altogether, these observations support that linoleic acid is oxidatively modified in the atherosclerotic aortic tissue. The detrimental role of linoleic acid in the progression of atherosclerosis is also supported by the studies on C57BL/6 mouse model fed for 15 weeks with an atherogenic diet containing conjugated linoleic acids (an isomer of linoleic acid) at 2.5-5 g/kg , resulting in increased development of the aortic fatty streaks [43].

Furthermore, Q-US/PS-NMF enabled to identify the signatures of oxidized all-*trans* retinol (cyan line) in the component spectra $\mathcal{C}_5$ (blue line), composed mainly of cholesterol (green line) and cholesteryl esters (magenta line) as shown in the inset of Fig. 5B. The appearance of a Raman band at 1529 cm^{-1} as well as a broad shoulder with peaks at 1618 cm^{-1} and 1635 cm^{-1} reflect oxidation-induced structural changes of all-*trans* retinol, ensuing in the process of its degradation [44], and therefore suggests its attribution to oxidized all-*trans* retinol. Importantly, the Raman feature at 1635 cm^{-1} in the $\mathcal{C}_5$ can be also assigned to the C=O stretch of ketones, previously observed in the spectrum of oxidized low density lipoprotein [45]. One more example of new information deduced from Q-US/PS-NMF is the presence of cholesterol in the β-carotene component spectrum ($\mathcal{C}_{11}$), which indicates that these species form a tightly-connected structure in the aortic tissue.

Notably, we also observed that $\mathcal{C}_9$ might be represented as a mixture of triglycerides and arachidonic acid at a ratio of about 3:1, but with a low R^2 value of 61%, which might indicate that these components are strongly modified in the atherosclerotic tissues due to oxidation.

To further claim we can quantify the information from the Raman images using Q-US/PS-NMF, we show in Fig. 5A-C the spatially-resolved Raman concentration maps of $\mathcal{C}_4$, $\mathcal{C}_5$, and $\mathcal{C}_8$ over the investigating images, representing atherosclerotic aortic tissue regions (A1, A2, A3) and non-atherosclerotic controls (C1, C2, C3). The similar figures of the remaining components are shown in Supplementary material. The concentration profiles of these components show that oxidized linoleic acid ($\mathcal{C}_8$) aggregates are strongly co-localized with cholesterol/saturated cholesteryl ester crystals ($\mathcal{C}_5$) in the atherosclerotic plaques as well as with oxidized lipid accumulations ($\mathcal{C}_4$) enriched with triglyceride and phosphocholine in the foam cells. This oxidized linoleic acid component shows one order of magnitude increase in concentration for the atherosclerotic plaque lesions compared to the atherosclerotic plaque-negative and non-atherosclerotic control regions. Altogether, this observation indicates the interplay between oxidized linoleic acid and cholesterol/triglyceride-rich components in the aortic tissue, which might be one of the possible molecular mechanisms involved in the evolution of the atherosclerotic plaque. Two independent studies, which used several different animal models of atherosclerosis, also found that oxidized fatty acids in the diet increase fatty streak lesion formation in the aortic tissues [46, 47]. Furthermore, our quantitative analysis reveals the significant levels of oxidative damage components strongly correlated with the atherosclerotic lesions in the aorta and therefore might indicate the persistent presence of inflammation known to be implicated in atherogenesis [48]. In particular, the components predominantly assigned to oxidized triglyceride ($\mathcal{C}_4$), cholesterol ($\mathcal{C}_5$) and linoleic acid ($\mathcal{C}_4$) and accumulated in the atherosclerotic lesions of the aortic tissues show about 20, 60, 12-fold increase in their maximum concentrations, respectively, compared to the non-atherosclerotic control regions (for comparison we use the samples A1-A3 and C3). The previous studies also support the oxidation hypothesis of atherogenesis, suggesting that the products of lipid peroxidation in the aorta modulate the innate immune system response, which triggers the uncontrolled development of foam cells, resulting in the initiation and progression of the atherosclerotic lesion [49, 50].

Also, the final results received from the quantitative Q-HIU analysis reveals the ability of the proposed method to detect both major and minor chemical components of atherosclerotic and non-atherosclerotic tissues. Table 2 shows that Q-HIU allows to identify low concentrated chemical compounds down to 0.04% from the original Raman data matrix characterised by fluorescent background of 94% and noise of 3%.

Altogether, we have demonstrated that the new developed Q-HIU method presented in this paper is capable to quantitatively decompose the spatially-resolved Raman spectra of human atherosclerotic aortic tissues into a number of individual biochemical components with relative mean Raman concentration down to 1%. These are structural proteins (actin, elastin, collagen), pathological lipid aggregates (cholesterol/cholesterol ester and triglyceride crystals), and minerals, which are characteristic biomarkers of atherosclerosis. Furthermore, our quantitative analysis of Raman micro-spectroscopy images allowed to identify and quantify oxidatively modified lipids (oxidative stress and inflammation bio-marker) found to be correlated/co-localized with the atherosclerotic lesions in the aorta, implying the use of the Q-HIU method in the quantitative characterization of human aortic tissues and potentially in the diagnosis of atherosclerosis.

Table 2: Relative mean Raman concentrations of noise, background and chemical components retrieved from the Q-HIU analysis of the biomedical Raman dataset of atherosclerotic tissues and non-atherosclerotic controls

	Normalization 1*		Normalization 2**	
	AS	non-AS	AS	non-AS
noise	3%	3%		
background	94%	93%		
$\mathcal{C}_1$	0.6%	3%	10%	35%
$\mathcal{C}_2$	0.7%	1%	13%	14%
$\mathcal{C}_3$	1%	0.6%	18%	8%
$\mathcal{C}_4$	1%	0.2%	19%	3%
$\mathcal{C}_5$	0.8%	0.3%	15%	4%
$\mathcal{C}_6$	0.5%	0.7%	9%	9%
$\mathcal{C}_7$	0.3%	0.7%	6%	10%
$\mathcal{C}_8$	0.6%	0.3%	10%	4%
$\mathcal{C}_9$	0.2%	0.2%	4%	2%
$\mathcal{C}_{10}$	0.09%	0.2%	2%	2%
$\mathcal{C}_{11}$	0.04%	0.1%	0.7%	2%

AS = atherosclerotic aortic tissue regions
* Relative mean Raman concentrations normalized on the mean original Raman intensity matrix (see Supplementary material for definition)
** Relative mean Raman concentrations of chemical components normalized according to Eqs. (18)-(19)

4.3.4. Comparison of Q-US/PS-NMF with existing methods

In this section, we will compare the performance of our Q-US/PS-NMF with the state-of-the-art methods such as MCR, FC-NMF and VCA using the same biomedical dataset of atherosclerotic aortic tissues as in the previous section. The results show that our realisation of NMF algorithm, Q-US/PS-NMF, which is implemented in a fast combinatorial ANLS framework, has exponential convergence to a local minimum with 200 and three times faster computational performance compared to MCR-ANLS [19] (see Fig. S1 of the Supplementary Material) and FC-NMF [21] (see Fig. S2 of the Supplementary material), respectively.

Comparison with the VCA method reveals that the relative factorization error in the 11 components' factorization analysis using this method was 11% versus 5% for our Q-US/PS-NMF. Also, the VCA was not able to identify cholesterol component, which is a known pathological biomarker of atherosclerosis, as well as linoleic acid. Furthermore, the R^2 values for the majority of component spectra found from the VCA are significantly lower compared to that for Q-US/PS-NMF (see Table 3 for detail). Overall, compared to VCA, Q-US/PS-NMF allows to identify additional chemical components which are characteristic bio-markers of the diseased tissues.

5. Conclusion

To conclude, we present a new efficient Quantitative Hyperspectral Image Unmixing (Q-HIU) method, allowing to autonomously analyse large-scale Raman microspectroscopy data with minimum input parameters (only two) and high accuracy (see an overview flowchart of Q-HIU analysis in Fig. 6). This in-house developed method integrates three consecutive steps of data analysis, called Singular Value Decomposition with innovative Automatic Divisive Correlation, Bottom Gaussian Fitting, and an efficient Quantitative Unsupervised/Partially Supervised Non-negative Matrix Factorization, which is capable to operate as a partially supervised method, when one or several samples' constituents are known. This is implemented by fixing the spectra of the known compounds in the approximating factorization expansion of the data matrix.

Compared with other existing hyperspectral image processing methods, Q-HIU shows significant improvement in accuracy as verified on both simulated and large volume Raman images of atherosclerotic tissues. The experimental results found from Q-HIU reveal a proof-of-principle for the biomolecular characterisation and quantitative imaging of individual biochemical components in the samples of various complexity (e.g.: the retrieval of individual component spectra and spatially-resolved concentration profiles of cholesterol/cholesteryl ester, oxidized linoleic acid, calcium hydroxyapatite, β-carotene crystals with relative mean Raman concentration down to 0.04% in atherosclerotic plaques compared to 0.1% in non-atherosclerotic controls from the original Raman intensity matrix with strong fluorescent background of 94% and noise of 3%). The great advantage of Q-HIU to directly work with matrices allows to achieve two orders of magnitude acceleration in computational speed as compared with the popular realization of MCR-ANLS.

Acknowledgements

The authors acknowledge S.G. Tikhodeev, N.A. Gippius, K. Triantafilou, P. Borri, W. Langbein, and F. Masia for useful discussions.

Funding

This work has been supported by the Cardiff University College of Biomedical and Life Sciences under the International Scholarship for PhD student and RFBR Project No. 16-29-03283.

Appendix A. Supplementary material

Supplementary material contains vector notations, definitions of Euclidean scalar product, correlation coefficient, and normalization on the mean original intensity matrix, as well as presents comparison of Q-US/PS-NMF

Table 3: Comparison of R^2 values for chemical components retrieved from Q-US/PS-NMF and VCA of the same Raman images as in Fig. 5

Chemical component attribution	R^2 for component number $\mathcal{C}_\#$		
	Q-US/PS-NMF with 11 components	VCA with 11 components	VCA with 9 components
actin	85% ($\mathcal{C}_1$)	77% ($\mathcal{C}_{10}$)	not found
water	94% ($\mathcal{C}_2$)	94% ($\mathcal{C}_2$)	not found
elastin	84% ($\mathcal{C}_3$)	90% ($\mathcal{C}_1$), 83% ($\mathcal{C}_9^*$)	89% ($\mathcal{C}_2$)
phospholipids + triglycerides	93% ($\mathcal{C}_4$)	91% ($\mathcal{C}_4$)	86% ($\mathcal{C}_3$)
cholesteryl esters	87% ($\mathcal{C}_5$)	91% ($\mathcal{C}_5$)	92% ($\mathcal{C}_6$)
+ vitamin A	89% ($\mathcal{C}_5$)	92% ($\mathcal{C}_5$)	93% ($\mathcal{C}_6$)
+ cholesterol	93% ($\mathcal{C}_5$)	not found	not found
collagen	83% ($\mathcal{C}_6$)	68% ($\mathcal{C}_3$), 92% ($\mathcal{C}_6^*$), 70% ($\mathcal{C}_8^*$)	91% ($\mathcal{C}_1$)
magnetite	75% ($\mathcal{C}_7$)	not found	not found
linoleic acid	90% ($\mathcal{C}_8$)	not found	not found
triglycerides + arachidonic acid	61% ($\mathcal{C}_9$)	not found	not found
hydroxyapatite	93% ($\mathcal{C}_{10}$)	66% ($\mathcal{C}_7$)	not found
β-carotene	80% ($\mathcal{C}_{11}$)	68% ($\mathcal{C}_{11}$)	68% ($\mathcal{C}_9$)
+ cholesterol	88% ($\mathcal{C}_{11}$)	not found	not found

Calculations of R^2 values were based on correlation of each component spectra with the pure biochemical references.
* indicates presence of water in the component.

with other software, convergence of Q-US/PS-NMF to local/global minima, results of the Q-HIU and VCA analysis, Raman spectra and concentration profiles of the dataset 2.

[1] M. Miljkovic, T. Chernenko, M. J. Romeo, B. Bird, C. Matthaus, M. Diem, Label-free imaging of human cells : algorithms for image reconstruction of Raman hyperspectral datasets, Analyst 135 (8) (2010) 2002–2013.

[2] C. Krafft, M. Schmitt, I. W. Schie, D. Cialla-May, C. Matthäus, T. Bocklitz, J. Popp, Label-Free Molecular Imaging of Biological Cells and Tissues by Linear and Nonlinear Raman Spectroscopic Approaches, Angew. Chem. Int. Ed. 56 (16) (2017) 4392–4430.

[3] L. Zhang, M. J. Henson, S. S. Sekulic, Multivariate data analysis for Raman imaging of a model pharmaceutical tablet, Anal. Chim. Acta 545 (2) (2005) 262–278.

[4] J. Filik, N. Stone, Analysis of human tear fluid by Raman spectroscopy, Anal. Chim. Acta 616 (2) (2008) 177–184.

[5] C. Krafft, I. W. Schie, T. Meyer, M. Schmitt, J. Popp, Developments in spontaneous and coherent Raman scattering microscopic imaging for biomedical applications, Chem. Soc. Rev. 45 (7) (2016) 1819–1849.

[6] K. Czamara, K. Majzner, A. Selmi, M. Baranska, Y. Ozaki, A. Kaczor, Unsaturated lipid bodies as a hallmark of inflammation studied by Raman 2D and 3D microscopy, Sci. Rep. 7 (18) (2017) 40889.

[7] R. Michael, A. Lenferink, G. F. J. M. Vrensen, E. Gelpi, R. I. Barraquer, C. Otto, Hyperspectral Raman imaging of neuritic plaques and neurofibrillary tangles in brain tissue from Alzheimer's disease patients, Sci. Rep. 7 (1) (2017) 15603.

[8] M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem, J. Popp, Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging, Theor. Chem. Acc. 130 (4-6) (2011) 1249–1260.

[9] J. J. Andrew, T. M. Hancewicz, Rapid analysis of Raman image data using two-way Multivariate Curve Resolution, Appl. Spectrosc. 52 (6) (1998) 797–807.

[10] S. Piqueras, L. Duponchel, R. Tauler, A. De Juan, Resolution and segmentation of hyperspectral biomedical images by Multivariate Curve Resolution-Alternating Least Squares, Anal. Chim. Acta 705 (2011) 182–192.

[11] B. Vajna, A. Farkas, H. Pataki, Z. Zsigmond, T. Igricz, G. Marosi, Testing the performance of pure spectrum resolution from Raman hyperspectral images of differently manufactured pharmaceutical tablets, Anal. Chim. Acta 712 (2012) 45–55.

[12] C. D. L. Albuquerque, R. J. Poppi, Detection of malathion in food peels by surface-enhanced Raman imaging spectroscopy and multivariate curve resolution, Anal. Chim. Acta 879 (2015) 24–33.

[13] C. H. Camp, M. T. Cicerone, Chemically sensitive bioimaging with coherent Raman scattering, Nat. Phot. 9 (5) (2015) 295–305.

[14] J.-X. Cheng, X. S. Xie, Vibrational spectroscopic imaging of living systems: an emerging platform for biology and medicine, Science 350 (6264) (2015) aaa8870.

[15] C.-Y. Lin, J. L. Suhalim, C. L. Nien, M. D. Miljkovic, M. Diem, J. V. Jester, E. O. Potma, Picosecond spectral coherent anti-Stokes Raman scattering imaging with principal component analysis of meibomian glands, J. Biomed. Opt. 16 (2) (2011) 021104.

[16] D. Fu, X. S. Xie, Reliable cell segmentation based on Spectral Phasor Analysis of hyperspectral Stimulated Raman Scattering imaging data, Anal. Chem. 86 (2014) 4115–4119.

[17] D. Zhang, P. Wang, M. N. Slipchenko, D. Ben-Amotz, A. M. Weiner, J.-X. Cheng, Quantitative vibrational imaging by hyperspectral Stimulated Raman Scattering microscopy and Multivariate Curve Resolution analysis, Anal. Chem. 85 (1) (2013) 98–106.

[18] C. Di Napoli, I. Pope, F. Masia, W. Langbein, P. Watson, P. Borri, Quantitative spatiotemporal chemical profiling of individual lipid droplets by hyperspectral CARS microscopy in living human adipose-derived stem cells, Anal. Chem. 88 (7) (2016) 3677–3685.

[19] J. Jaumot, A. D. Juan, R. Tauler, MCR-ALS GUI 2.0 : New features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12.

[20] M. H. Van Benthem, M. R. Keenan, Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems, J. Chemom. 18 (10) (2004) 441–450.

[21] H. Kim, H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, Bioinformatics 23 (12) (2007) 1495–1502.

[22] C. Lawson, R. Hanson, Solving Least Squares Problems, SIAM,

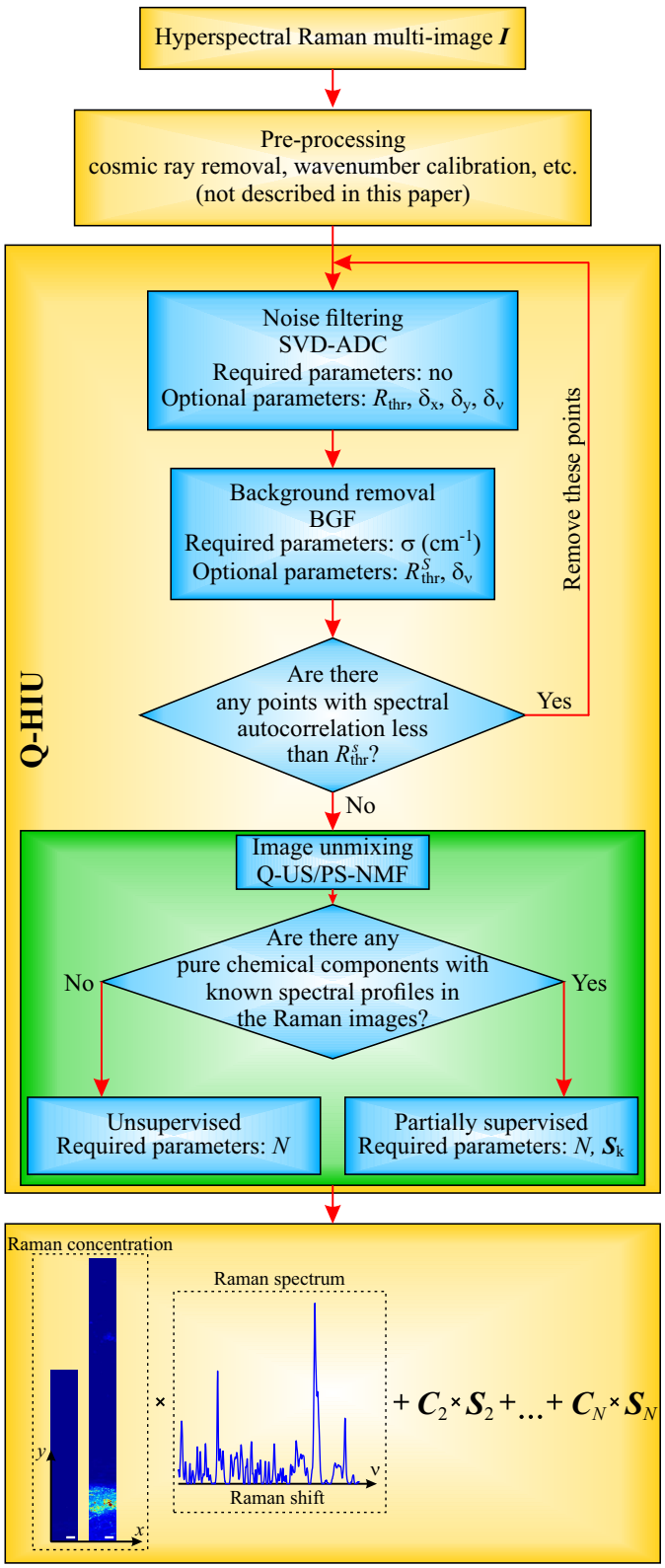

Figure 6: Flowchart of Q-HIU analysis.

Philadelphia, 1995.

[23] G. H. Golub, C. F. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore, 2013.

[24] N. Uzunbajakava, A. Lenferink, Y. M. Kraan, E. Volokhina, G. Vrensen, J. Greve, C. Otto, Nonresonant confocal Raman imaging of DNA and protein distribution in apoptotic cells, Biophys. J. 84 (6) (2003) 3968–81.

[25] K. Hirokawa, Aids for analytical chemists, Anal. Chem. 52 (12) (1980) 1966–1968.

[26] S.-J. Baek, A. Park, J. Kim, A. Shen, J. Hu, A simple background elimination method for Raman spectra, Chemom. Intell. Lab. Syst. 98 (1) (2009) 24–30.

[27] P. H. Eilers, A perfect smoother, Anal. Chem. 75 (14) (2003) 3631–3636.

[28] P. H. Eilers, Parametric time warping, Anal. Chem. 76 (2) (2004) 404–411.

[29] A. Jirasek, G. Schulze, M. M. L. Yu, M. W. Blades, R. F. B. Turner, Accuracy and precision of manual baseline determination, Appl. Spectrosc. 58 (12) (2004) 1488–1499.

[30] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. De Juan, A. Gorzsás, Vibrational spectroscopic image analysis of biological material using multivariate curve resolution-alternating least squares (MCR-ALS), Nature Protocols 10 (2) (2015) 217–240.

[31] P. Paatero, U. Tapper, Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, Environmetrics 5 (2) (1994) 111–126.

[32] P. Paatero, Least squares formulation of robust non-negative factor analysis, Chemometr. Intell. Lab. Syst. 37 (1) (1997) 23–35.

[33] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–91.

[34] D. D. Lee, H. S. Seung, Algorithms for Non-negative Matrix Factorization, in: Adv. Neural Inf. Process Syst. Vol. 13, MIT Press, Massachusetts, 2001, pp. 556–562.

[35] A. Y. F. You, M. S. Bergholt, J.-P. St-Pierre, W. Kit-Anan, I. J. Pence, A. H. Chester, M. H. Yacoub, S. Bertazzo, M. M. Stevens, Raman spectroscopy imaging reveals interplay between atherosclerosis and medial calcification in the human aorta, Sci. Adv. 3 (12) (2017) e1701156.

[36] A. A. Green, M. Berman, P. Switzer, M. D. Craig, A transformation for ordering multispectral data in terms of image quality with implications for noise removal, IEEE Trans. Geosci. Remote Sens. 26 (1) (1988) 65–74.

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

[38] J. M. P. Nascimento, J. M. B. Dias, Vertex component analysis: A fast algorithm to unmix hyperspectral data, IEEE Transactions on Geoscience and Remote Sensing 43 (4) (2005) 898–910.

[39] A. Y. F. You, M. S. Bergholt, J.-P. St-Pierre, W. Kit-Anan, I. J. Pence, A. H. Chester, M. H. Yacoub, S. Bertazzo, M. M. Stevens, Research data supporting "Raman spectroscopy imaging reveals interplay between atherosclerosis and medial calcification in human aorta" (2017).

[40] E. G. Lobanova, S. V. Lobanov, K. Triantafilou, W. Langbein, P. Borri, Quantitative chemical imaging of amyloid-β plaques with Raman micro-spectroscopy in human Alzheimer's diseased brains, arXiv:1803.01201arXiv:1803.01201.

[41] K. Czamara, K. Majzner, M. Z. Pacia, K. Kochan, A. Kaczor, M. Baranska, Raman spectroscopy of lipids: A review, J. Raman. Spectrosc. 46 (1) (2015) 4–20.

[42] C. Krafft, L. Neudert, T. Simat, R. Salzer, Near infrared Raman spectra of human brain lipids, Spectrochim. Acta A 61 (7) (2005) 1529–1535.

[43] J. S. Munday, K. G. Thompson, K. A. James, Dietary conjugated linoleic acids promote fatty streak formation in the C57BL/6 mouse atherosclerosis model, Br. J. Nutr. 81 (3) (1999) 251–255.

[44] N. Failloux, I. Bonnet, M.-H. Baron, E. Perrier, Quantitative analysis of vitamin A degradation by Raman spectroscopy, Appl. Spectrosc. 57 (9) (2003) 1117–1122.

[45] D.-S. Wang, N. Iwata, E. Hama, T. C. Saido, D. W. Dickson, Oxidized neprilysin in aging and Alzheimer's disease brains, Biochem. Biophys. Res. Commun. 310 (1) (2003) 236–241.
[46] I. Staprans, X. M. Pan, J. H. Rapp, K. R. Feingold, The role of dietary oxidized cholesterol and oxidized fatty acids in the development of atherosclerosis, Mol. Nutr. Food Res. 49 (11) (2005) 1075–1082.
[47] N. Khan-Merchant, M. Penumetcha, O. Meilhac, S. Parthasarathy, Oxidized fatty acids promote atherosclerosis only in the presence of dietary cholesterol in low-density lipoprotein receptor knockout mice, J. Nutr. 132 (2002) 3256–3262.
[48] P. Libby, Inflammation in atherosclerosis, Nature 420 (6917) (2002) 868–874.
[49] J. S. Bus, J. E. Gibson, Lipid peroxidation and its role in atherosclerosis, Br. Med. Bull. 49 (3) (1993) 566–576.
[50] M. Navab, G. M. Ananthramaiah, S. T. Reddy, B. J. Van Lenten, B. J. Ansell, G. C. Fonarow, K. Vahabzadeh, S. Hama, G. Hough, N. Kamranpour, J. A. Berliner, A. J. Lusis, A. M. Fogelman, The oxidation hypothesis of atherogenesis: the role of oxidized phospholipids and HDL, J. Lipid Res. 45 (6) (2004) 993–1007.