*Submitted Paper*

Applied Spectroscopy
www.s-a-s.org

Applied Spectroscopy
1–10
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00037028241311296
journals.sagepub.com/home/asp
Sage

# Demixing and Analysis of Complex Biological Raman Hyperspectra Based on Peak Fitting, Amplitude Trend Clustering, and Spectrum Reconstruction

**H. Georg Schulze[1], Shreyas Rangan[2,3], Martha Z. Vardaki[4], Michael W. Blades[5], Robin F. B. Turner[2,5,6], and James M. Piret[2,3,7]**

## Abstract

To better interpret the Raman spectra from mammalian cells, it is often desirable to reduce their complexity by decomposing them into the spectral contributions from individual macromolecules or types of macromolecules. Diverse methods exist for demixing complex spectra, each with different benefits and drawbacks. However, some methods require a library of component spectra that might not be available, while others are hampered by noise and peak congestion that includes many proximal overlapping peaks. Through rapid fitting of individual peaks in every spectrum of a Raman hyperspectral data set, we have obtained individual peak parameters from which we determined the trends for all the peak amplitudes. We then grouped similar trends with *k*-means clustering. Then we used the peak parameters of all the peaks in a given cluster to reconstruct a spectrum representative of that cluster. This method produced spectra that were less distorted by unrelated overlapping peaks or noise, were less congested than those in the hyperspectral set, and thereby improved peak identification and macromolecule recognition. We have demonstrated the application of the method with Raman spectra from a perchlorate–polystyrene model system and extended it to complex spectra from methanol-fixed mammalian cells. We were able to recover independent spectra of perchlorate and polystyrene in the model system and spectra pertaining to individual macromolecular types (proteins, nucleic acids, lipids) from the mammalian cell data. We discuss how imperfections in spectral preprocessing and peak fitting can adversely affect the results. In summary, we have provided a proof-of-concept for a novel mixture resolution method with different attributes than extant ones.

## Keywords

Raman spectroscopy, *k*-means clustering, peak fitting, mixture resolution, mammalian cell analysis, hyperspectral analysis

Date received: 11 September 2024; accepted: 9 December 2024

[1]Independent, Monte do Tojal, Hortinhas, Terena, Portugal
[2]Michael Smith Laboratories, The University of British Columbia, Vancouver, British Columbia, Canada
[3]School of Biomedical Engineering, The University of British Columbia, Vancouver, British Columbia, Canada
[4]Institute of Chemical Biology, National Hellenic Research Foundation, Athens, Greece
[5]Department of Chemistry, The University of British Columbia, Vancouver, British Columbia, Canada
[6]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, British Columbia, Canada
[7]Department of Chemical and Biological Engineering, The University of British Columbia, Vancouver, British Columbia, Canada

**Corresponding Authors:**
Robin F. B. Turner, Michael Smith Laboratories, The University of British Columbia, 2185 East Mall, Vancouver, British Columbia, V6T 1Z4, Canada; Department of Chemistry, The University of British Columbia, 2036 Main Mall, Vancouver, British Columbia, V6T 1Z1, Canada; Department of Electrical and Computer Engineering, The University of British Columbia, 2332 Main Mall, Vancouver, British Columbia, V6T 1Z4, Canada.
Email: turner@msl.ubc.ca

James M. Piret, Michael Smith Laboratories, The University of British Columbia, 2185 East Mall, Vancouver, British Columbia, V6T 1Z4, Canada; School of Biomedical Engineering, The University of British Columbia, 2222 Health Sciences Mall, Vancouver, British Columbia, V6T 1Z3, Canada; Department of Chemical and Biological Engineering, The University of British Columbia, 2360 East Mall, Vancouver, British Columbia, V6T 1Z3, Canada.
Email: james.piret@ubc.ca

## Introduction

In carrying out their functions within various tissues, mammalian cells employ intricate and dynamic metabolic processes that can change their chemical makeup. Measuring these transient or permanent changes in chemical composition provides a means to make inferences about many aspects of the changing nature and condition of cells. This is particularly important when there is a need to effectively and efficiently manipulate cells and their biological processes to manufacture reliably safe and efficacious biological therapeutics.[1–3] However, the rapid and accurate determination of the chemical composition of cells is essential to integrate this information into manufacturing process monitoring, control, and validation.

Raman spectroscopy is an information-rich vibrational spectroscopic method with unique attributes that make it well-suited for use in biology and medicine. Importantly, it can be nonperturbing and suitable for use with live cells in a label-free manner, such that it requires little or no sample preparation. It can be performed rapidly and remotely with optical fibers or combined with microscopy.[4–8] Despite being based on a weak optical scattering process,[4] major laser and detector technological advances have enabled the development of Raman techniques that permit the rapid acquisition of high-quality spectra, that to a considerable extent, can be quantitative.[8] These advances also have been enabled by improved spectral preprocessing and analysis techniques, many of which are fully automated to produce rapid results without introducing user bias.[9,10]

These attributes have resulted in Raman spectroscopy already possessing a proven track record in biological and biomedical research. This includes assessments of the stages of the normal cell cycle chromosomal duplication,[11,12] radiation-induced metabolic changes in cancer cells,[13] distinguishing types of cell death,[14] the stages of cell differentiation to types with increasingly specialized functions,[5] the stimulation of cells to secrete specific hormones,[15] and the transformation of immune cells from one type to another.[16]

Due to the complexity of spectra with a multitude of overlapping bands coming from the many molecular species present in cells, the quantitative and qualitative analyses of spectra are challenging. Though inferences and analyses can be based on one or a small subset of peaks that are indicative of a given macromolecule, it is desirable to identify complete[17] or highly similar[18] spectra of a given macromolecule or macromolecular type.[19] Several spectral multivariate demixing methods can be deployed to attain such aims such as principal component analysis (PCA),[20] two-dimensional correlation spectroscopy (2D-COS),[21] non-negative least squares using a library of known spectra, or non-negative matrix factorization (NNMF).[13,22] Each of these has drawbacks that can render their use problematic. Correlation-based methods such as PCA and 2D-COS suffer from comingling of correlated and anticorrelated changes that can misrepresent the correlated behavior of overlapping peaks. Library-based methods require the use of a suitable extant library and NNMF can be affected by noise in the spectra and obtaining good starting values for the factorization.[23,24]

We report here another method which exhibits different advantages and drawbacks and so complements those in current use. We have previously developed an algorithm to rapidly decompose all spectra in a Raman hyperspectral set by fitting individual peaks to all the spectra in the set.[25] Fitting proper probability distributions to individual peaks in a spectrum resolves overlapping peaks into their components and so establishes a better approximation to the true number of peaks as well as their correlation structure in a Raman hyperspectral set. We have also developed an algorithm to perform *k*-means clustering on the intensity trends of the peaks in hyperspectra.[26] This method sorts the intensity changes along the individual wavenumbers of a Raman hyperspectral set into separate groups. Changes within groups are similar, and changes between groups are qualitatively different even though possibly correlated. The approaches are here combined into a method that proceeds by fitting peaks to all the spectra in a set of hyperspectra, grouping together all the intensities of the fitted peaks that change in a similarly patterned manner, and reconstructing representative spectra for each group from the fit parameters of all the peaks in that group. We used a simple polystyrene–perchlorate model system exhibiting clear and contrasting peak trends to demonstrate the basic concept and then tested it with a highly complex system consisting of spectra obtained from methanol-fixed mammalian cells that exhibited more convoluted trends.

## Experimental

### Materials and Methods

*Vector-Based Peak Fitting.* The reader is referred to our previous work for complete details of the vector-based peak fitting method using an alternating least squares procedure.[25] Briefly, the first part of this two-part algorithm consists of a moving window peak fitting procedure to provide an initial estimate of the number of peaks present in the spectra and their parameters. We use these estimated parameters in part two in which we approximate Raman peaks with the Gaussian distribution:

$$\mathbf{intensities} = \left(\frac{\mathbf{amplitudes}}{\mathbf{sigmas} \times \sqrt{2\pi}}\right) e^{-0.5\left(\frac{\mathbf{x-positions}}{\mathbf{sigmas}}\right)^2} \quad (1)$$

where **x** is a vector of spectral locations selected to simplify Eq. 1, **intensities** and **amplitudes** are vectors of spectral intensities and peak amplitudes, respectively, corresponding to the locations in **x**, **positions** is a vector of band positions and **sigmas** a vector of band standard deviations. We exploited features of this simplified equation to calculate peak sigmas, amplitudes, and positions for all the spectra in the dataset using the parameters estimated in the first part. These parameters were calculated iteratively, and the results of each iteration were then used in the next iteration until a preset threshold for the maximum number of iterations was reached. The parameters obtained for the minimum root mean square error between the sum of the Gaussian distributions and the spectrum being fitted were retained. For this approach, we approximate Raman peaks with Gaussian distributions instead of the more appropriate Voigt distributions due to the former being mathematically more tractable.

*K-Means Clustering of Fitted Peak Amplitudes.* *K*-means clustering is an unsupervised grouping procedure[27–29] where the number of clusters, *k*, must be specified at the outset. The trend from the amplitudes of every peak, as obtained from peak fitting, is then assigned to exactly one of *k* nonempty clusters by an iterative procedure. Initial cluster centroids are randomly assigned from the data set, and the distance between every trend and every centroid is determined using a chosen distance measure. Every trend is then assigned to the cluster of its nearest centroid, and every centroid is then modified to be the mean trend of all its cluster members. Therefore, the distances between every trend and every centroid might now be different and they are recalculated. Some trends might now also be closer to another centroid, and they are reassigned to that cluster. The resultant changes in cluster members produce further centroid modifications with new mean trends. When no further cluster changes occur or a specified iteration number is reached, the process is terminated. Drawbacks include algorithm convergence to a local minimum and the need for the user to choose an appropriate number of clusters for a given task.

*Model System.* We used a model system, consisting of an aqueous 2.5 M sodium perchlorate ($NaClO_4$) solution in a Petri dish wherein was submerged a 12.5-mm diameter glass-encapsulated gold mirror (ThorLabs, USA) with a monolayer of ~10 µm diameter polystyrene microspheres (Sigma-Aldrich, USA) deposited on it, to test the combination of peak fitting and peak trends clustering algorithms.

*Cell Culture and Fixation.* Approximately $2 \times 10^6$ human Jurkat T-cells (ATCC, TIB-152) were seeded into T-75 flasks containing 15 mL Immunocult-XF medium (Stemcell Technologies, Canada) supplemented with 1× antibiotic–antimycotic cocktail (GIBCO, Grand Island, NY). They were then incubated in a humidified incubator at 37 °C and 5% $CO_2$ for 72 h. Exponentially growing cells were harvested and fixed for Raman spectroscopy. Different volumes (12.5, 25, 37.5, or 50 µL) of methanol, augmented with water to 50 µL, were used to perform fixation in 25, 50, 75, or 100% methanol on four groups of approximately $2 \times 10^6$ collected and centrifuged Jurkat cells. After removing the supernatant and washing the cells once with saline, the cell pellets were resuspended in one of the given percentages of methanol and incubated at –20 °C for 20 min. Then the cell/methanol suspensions were pipetted onto 12.5 mm diameter glass-encapsulated gold mirrors (ThorLabs, USA), air-dried in a biosafety cabinet, and thereafter stored at 4 °C for Raman spectroscopy.

*Raman Spectroscopy of the Model System.* We scanned across the edge of a close-packed area of beads where a transition from a bead-free to a bead-filled area occurred as shown in Figure 1a. The scanning produced decreasing perchlorate ion and increasing polystyrene peak intensities, thus producing opposite trends. The major perchlorate ion Raman band at ~935 $cm^{-1}$ is distinct from the strong polystyrene Raman bands as evident in Figure 1b. The Petri dish containing the mirror with beads was placed under a confocal Raman microscope (InVia, Renishaw, Gloucestershire, UK), and a water-immersion 40× (0.80 NA, 3300 µm working distance, Leica Microsystems, Germany) objective lens was used to collect spectra from a laser beam focal spot of approximately 3 µm × 30 µm. Thirty-seven Raman spectra, one per ~1 µm step while stepping through the edge of the bead array, were collected. Spectra were collected for 1 s each at room temperature and with a 50 µm slit width, 785 nm laser excitation, and ~80 mW power at the sample.

*Raman Spectroscopy of Fixed Cells.* From 50 to 60, Raman spectra were recorded in map acquisition mode from each of the four groups of fixed cell samples in two biological replicates for a total of 431 spectra. We used a Raman microscope (Renishaw, inVia, UK) with a 785 nm diode laser and a 50× long focal length objective lens (Leica Microsystems, Germany). The acquisition time per spectrum was 10 s, and each spectrum contained information from 10 to 15 cells. These spectra are shown in Figure 1c. A variety of literature sources were consulted for Raman band assignments.[30,31]

*Data Generation and Processing.* Matlab R2017b (The MathWorks Inc., USA) was used for spectral processing and data analyses. The Raman spectra were preprocessed in a previously described integrated manner[9,10] consisting of an automated moving average-based baseline-flattening method (15 iterations),[32] a two-dimensional second difference cosmic ray-induced spike removal method[33] and then a contiguous single-channel Voigt distribution fitting method for smoothing.[34]

Peak parameters were estimated from a "starting" spectrum by moving a seven-peak window across the mean spectrum of the constant-sum normalized spectra of a series and fitting Gaussian distributions simultaneously to all the peaks in the window as described previously.[35] The peak parameters of all the peaks in every spectrum were then rapidly

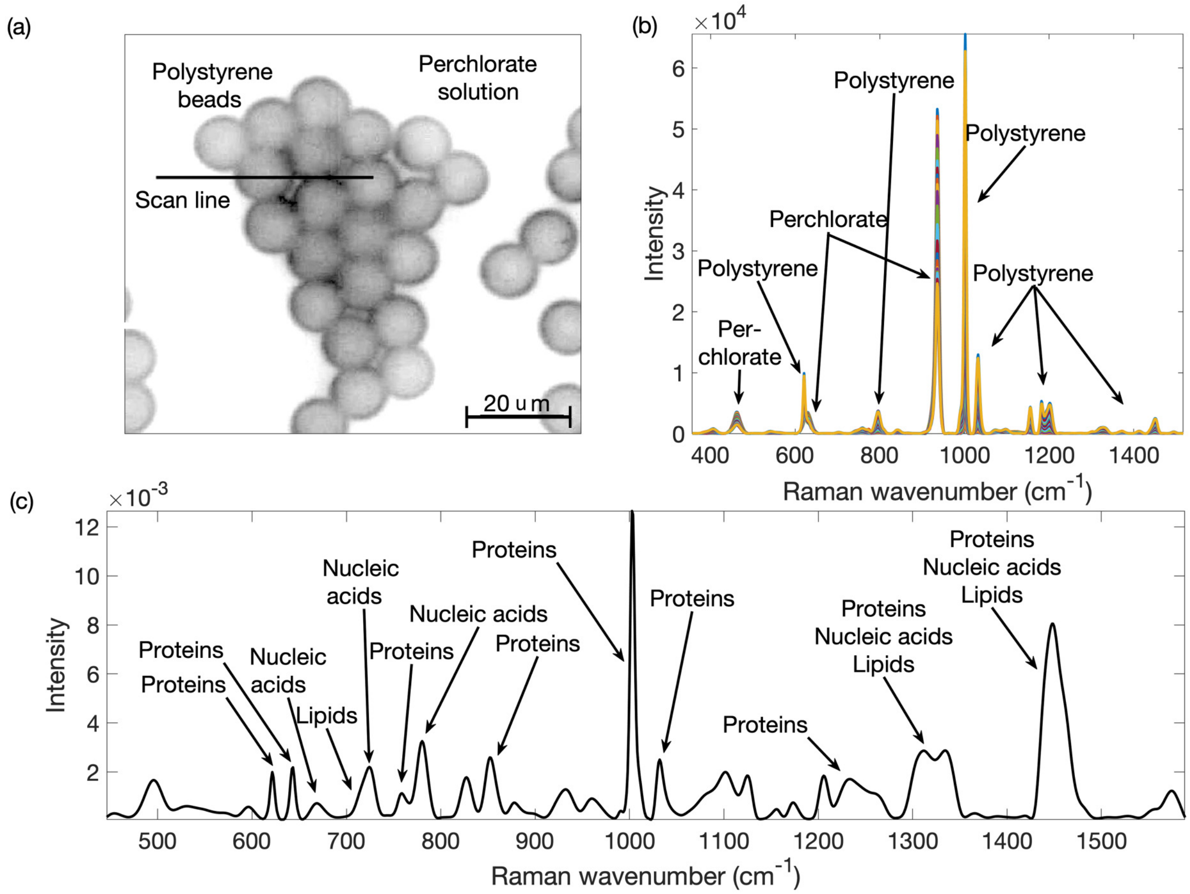

**Figure 1.** (a) Image of polystyrene-perchlorate model system showing the line scan performed from left to right. (b) Thirty-eight Raman spectra from the model system; spectra were not normalized. (c) Mean of constant sum normalized Raman spectra from Jurkat cells fixed with different percentages of methanol.

calculated with an iterative alternating least squares method.[25] Two separate momentum terms were used to allow moderation of the size of the alternating least squares adjustments, that is, one for the peak positions and the other for the remaining peak parameters.

For the model system, estimating initial peak parameters from the mean "starting spectrum", peak positions were allowed to shift by ~15 wavenumbers to either side of the initial position estimate, their maximum amplitudes allowed to vary up to 100 times that of the maximum peak intensity in the series of measured spectra, and a spectral resolution (peak width at half maximum) of 8 cm$^{-1}$ was used with peak widths that could vary up to three times that much. Subsequently, when calculating the peak parameters for every spectrum in the series with the vector-based method, we used 100 iterations, a 6 cm$^{-1}$ spectral resolution, and a momentum term of 0.001 for adjusting calculated parameters on each iteration except for peak positions that were held fixed.

When estimating peak parameters for the Jurkat cell spectra, peak positions were allowed to shift only by one wavenumber to either side of the estimated position because the spectra contained many overlapping peaks. The maximum amplitudes were allowed to vary up to 100 times that of the maximum peak intensity, and a spectral resolution of

4 cm$^{-1}$ was used with peak widths that could vary up to three times that much. When thereafter calculating the peak parameters for all the spectra in the series, we used 15 iterations, a 4 cm$^{-1}$ spectral resolution, and a momentum term of 0.001 for adjusting calculated parameters on each iteration except for peak positions that were held fixed by using a peak position momentum of 0.

These parameters were determined based on the nature of the spectra. For spectra with sparse peaks, more liberal allowances in peak shifts can be made, but for spectra with many overlapping peaks, shifts need to be constrained to prevent convergence to inappropriate local minima. Similar considerations applied to peak width selection and how much the widths could reasonably vary given the spectra being processed. A generous allowance was made for variations in peak amplitudes because they were not considered to be

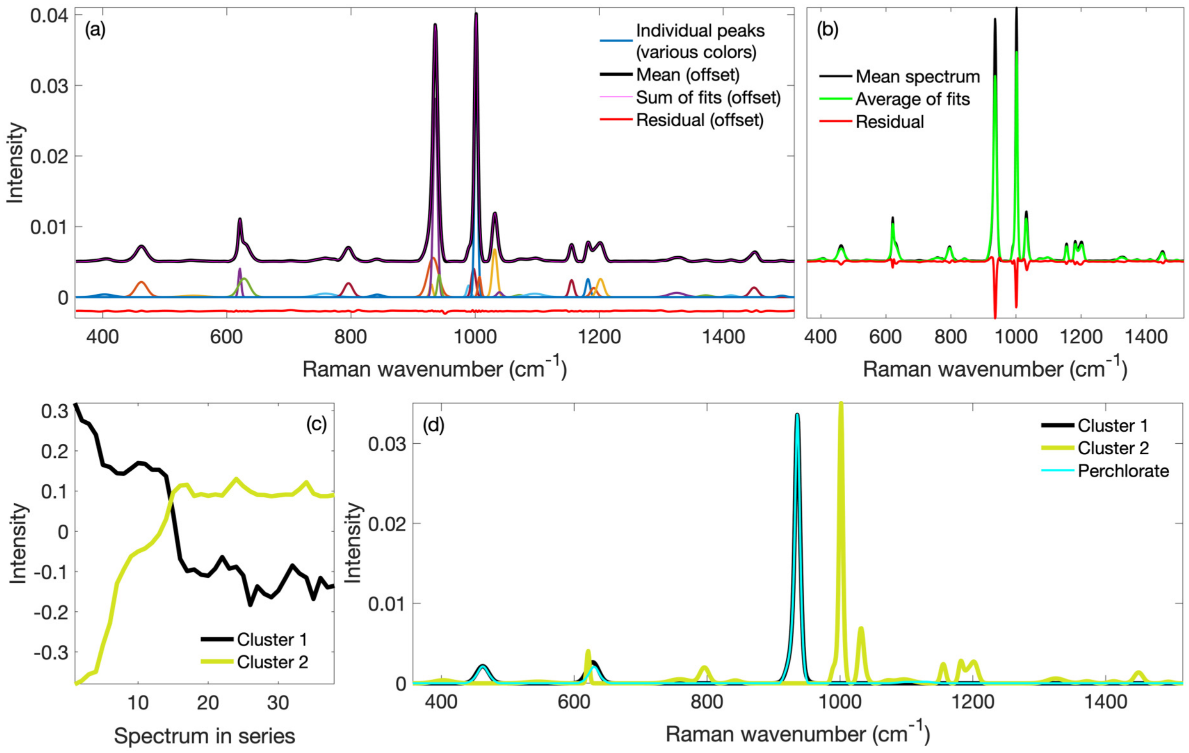

**Figure 2.** (a) The individual fitted bands obtained from the mean of the constant sum-normalized spectra recorded from the model system are shown in various colors. Their sum is shown in magenta superimposed on the black mean spectrum (both with positive offset for ease of viewing); also shown, with negative offset, is the residual (red) between the mean spectrum and the sum of the fitted bands. Band parameters determined from the fits to the mean spectrum peaks were used as initial values to calculate band parameters for all the peaks in a given spectrum, and this procedure was repeated for all the spectra. (b) These latter band parameters were then used to calculate an estimate or fit of that spectrum. Shown is the average of the fitted spectra (green) overlain on the mean spectrum (black) in the set along with the residual difference between them (red). (c) Standard normal variate trends of the intensities obtained from the peak fits were separated into two clusters with *k*-means clustering. From the individual peaks sorted into a given cluster, a spectrum was generated. These cluster-based spectra are shown in (d) with a superimposed perchlorate spectrum. The superposition identifies the Cluster 1 spectrum to be that of perchlorate and demonstrates the mixture resolution nature of this approach. The *y*-axis label of (a) also pertains to (b).

of critical importance once peak positions, and peak widths were established.

The Matlab *kmeans* function, with the correlation distance measure, was used for *k*-means clustering of the standard normal variate-normalized trends of every calculated peak intensity. For the model system, we used two clusters to represent perchlorate and polystyrene; for the Jurkat cell spectra, we used five clusters to represent the major cellular macromolecule types (proteins, nucleic acids, lipids, carbohydrates, and others). The peak parameters of all the peaks within a given cluster were then used to reconstruct a spectrum.

The Matlab code, which requires Matlab R2017b (but does not work on later releases), is available from the authors upon request.

## Results and Discussion

The starting spectrum for obtaining initial peak parameters for the hyperspectral data set from the model system, along with the individual peaks generated from those parameters, is shown in Figure 2a. These estimates were used to calculate peak parameters for every peak in a spectrum and from these parameters, a Gaussian distribution for every peak was generated. All distributions were then added to provide an approximation of that spectrum. The average of the approximated spectra is shown as a green trace in Figure 2b that is superimposed on the black mean spectrum of the series.

In Figure 2a, it can be seen that the slight asymmetry of the major perchlorate and polystyrene bands resulted in four overlapping peaks fitted to each of them. These produced summations with near-perfect fits to the original bands as can be determined from the residual. Though the rapid fitting of the individual spectra in the series produced deviations as shown in Figure 2b residual, all the peaks pertaining to a given component grouped together and their summations resulted in the spectra of the pure components as explained next.

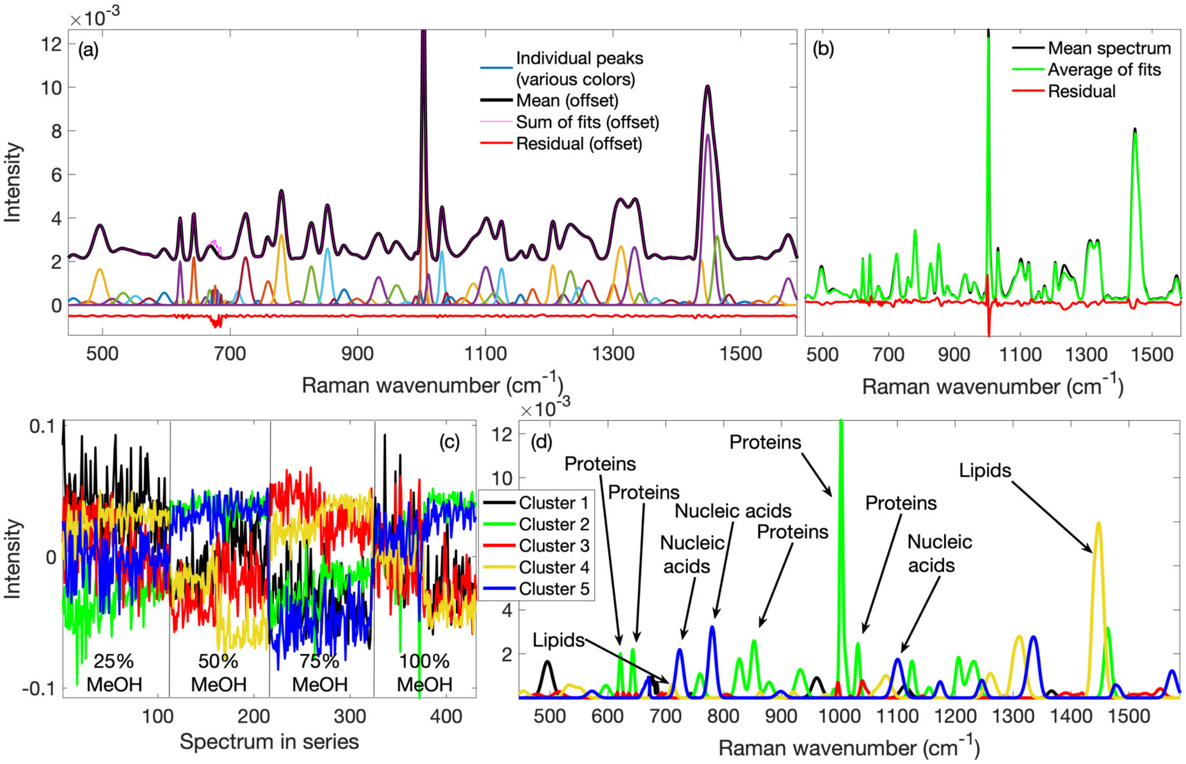

**Figure 3.** (a) The individual fitted bands obtained from the mean of the constant sum-normalized spectra are shown in various colors. Their sum is shown in magenta superimposed on the black mean spectrum (both with positive offset for clarity); also shown is the red residual between mean and sum of fitted spectra (with negative offset). Band parameters determined from the fits to the mean spectrum's peaks were used as initial values to calculate band parameters for all the peaks in every spectrum of the set. (b) Band parameters determined for all the peaks in a given spectrum were used to calculate a fit or estimate of that spectrum. Shown is the average of all the fitted spectra overlain on the mean spectrum in the set along with the difference between them. (c) Standard normal variate trends of the peaks were separated into five clusters with *k*-means clustering. From the individual peaks sorted into a given cluster, a spectrum was generated. The cluster-based spectra are shown in (d) and show that protein-related peaks, nucleic acid–related peaks, and lipid-related peaks were sorted into separate clusters. The *y*-axis label of (a) also pertains to (b) and the cluster legend pertains to both (c) and (d).

Standard normal variate trends of the intensities obtained from fits to every peak across all spectra were separated into two clusters with *k*-means clustering. The trends are shown in Figure 2c. From those individual fitted peaks in the starting spectrum that were sorted into a given cluster, a single spectrum was generated. The cluster-based generated spectra are shown in Figure 2d with a superimposed perchlorate spectrum scaled to equal maximum Cluster 1 peak height. The excellent agreement between the superimposed perchlorate spectrum and the generated Cluster 1 spectrum identifies Cluster 1 spectrum as that of perchlorate and Cluster 2 as that of polystyrene. It also demonstrates that mixture resolution can be affected with this approach. In principle, the number of mixture components that can be digitally demixed is expected to depend on the number of components with differential trends that are not compromised by noise.

The starting spectrum for obtaining initial peak parameters for the hyperspectral data set from Jurkat cells fixed with various percentages of methanol was the constant sum-normalized mean spectrum of the set. Its resolution through fitting 88 individual peaks is shown in Figure 3a with the difference between their sum and the starting mean spectrum, offset for ease of viewing, attesting to an effective fitting. The sum of the rapid fits of all individual spectra in the hyperspectral set is shown as a green trace in Figure 3b. It is mostly superimposed on the black starting spectrum that was the mean of all the constant sum normalized hyperspectra. Their difference is shown as a red trace. The standard normal variate trends are shown in Figure 3c. From those individual fitted peaks that were sorted into a given cluster, a spectrum was generated. The cluster-based generated spectra are shown in Figure 3d, where it can be seen that peaks belonging to proteins, nucleic acids, and lipids were sorted into separate clusters.

In Figure 4, we show separately the cluster-based spectra of Figure 3d with detailed band assignments based on published libraries.[30,31,36] From these detailed band assignments, their macromolecular natures were confirmed. The spectrum reconstructed from Cluster 2 contained only protein peaks, that from Cluster 5 only nucleic acids peaks and

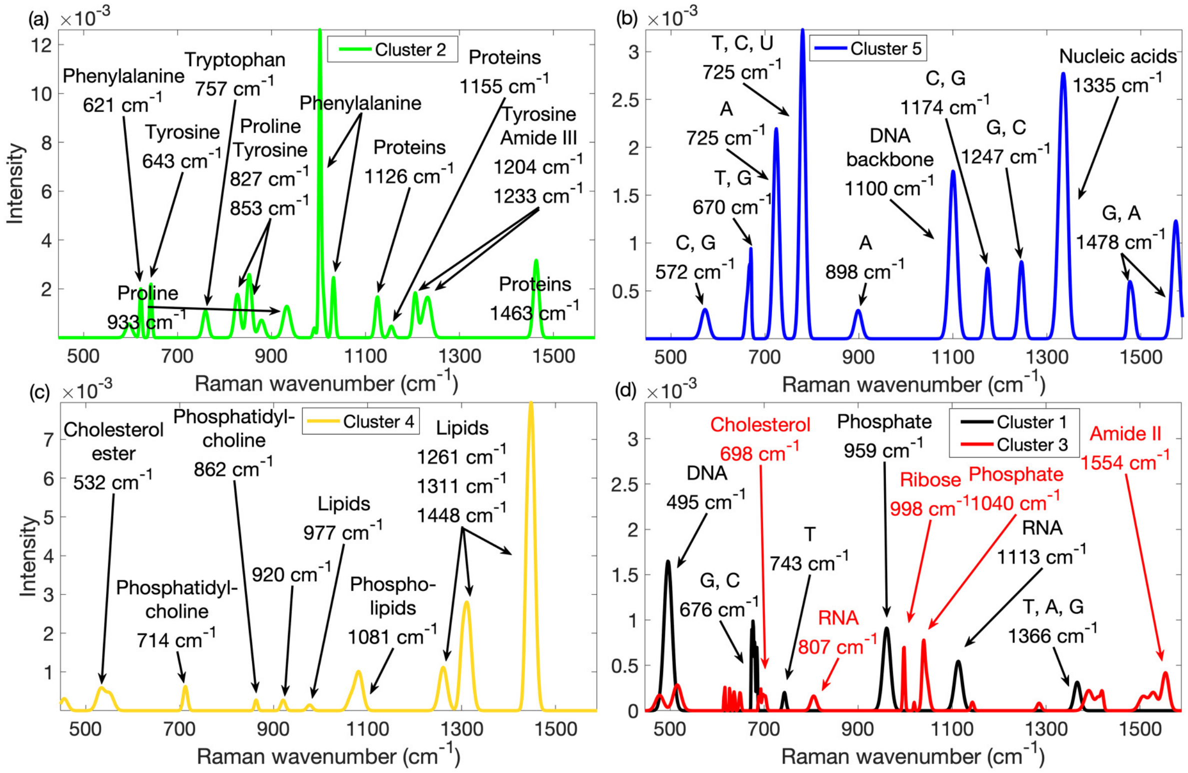

**Figure 4.** All the peaks in the Jurkat cell spectrum reconstructed from the fit parameters of the peaks sorted, based on their standard normal variate trends, into (a) Cluster 2 were all exclusively related to proteins as is evident from the detailed assignments; (b) those in Cluster 5 were exclusively related to nucleic acids as is evident from the detailed assignments and (c) the detailed assignments of those in Cluster 4 were related to lipids. (d) No cluster consisting of carbohydrates was obtained. Instead, Cluster 1 spectrum peaks could also be assigned to nucleic acids. Cluster 3 peaks could not consistently be assigned and seemed to belong to proteins, nucleic acids, and lipids. The *y*-axis label of (a) also pertains to (b) and that of (c) to (d).

that from Cluster 4 only lipid peaks. We did not observe any noticeable evidence of carbohydrates in the remaining two clusters. Instead, the peaks on Cluster 1 could also be assigned to nucleic acids. The spectrum from Cluster 3 had the weakest peaks of all the clusters and consistent assignments could not be performed as the peaks appeared to come from all three macromolecular types mentioned above.

Figure 4 demonstrates that, by using this methodology, most of the biocomponent or macromolecular type spectra can be reconstructed from Raman hyperspectra. An advantage of this approach is that, in principle, the reconstructed spectra from fitted peaks might produce the underlying spectra of cell components as they exist in the sample. Clustering also caused peaks to be partitioned into exclusive groups, thus cluster-based spectra are less congested and have fewer overlapping peaks as can be seen in a comparison of Figure 3a with any of the panels in Figure 4. These effects should ease peak identification by reducing the overlap between peaks.

We did not scan the samples with a laser spot size small enough to effect intracellular resolution. Instead, the laser spot size tended to include three to four cells. Though this might have caused some variation in the relative abundances of macromolecules present within the laser spot when taking a spectrum, it is likely that the methanol treatments would have induced more substantial variations in the probed samples. If this were so, Figure 4d suggested that fixing with methanol differentially affected proteins, nucleic acids, and lipids because peaks from these macromolecules clustered together. Differential methanol fixation effects might also have occurred within the nucleic acid group because, besides Cluster 5, a subset of nucleic acid peaks were assigned to Cluster 1. Thus, a separate assignment might have occurred based on fixative-induced changes to different nucleic acids or different structural parts of the nucleic acids.[37,38] Clustering might therefore facilitate the interpretation of experimental effects. Another attribute of the methodology is that cluster-based reconstruction uses peak parameters obtained from peak fitting, thus such spectra are devoid of noise.

There are also several issues that limit the utility of this approach. We have opted to use here five clusters representing the major cellular macromolecule types (proteins, nucleic acids, lipids, carbohydrates, and others) because of their anticipated relative abundances and the potential utility of such demixing. Literature reports indicate that these types of major macromolecules constitute up to 95% of cellular dry mass.[39–44] This also constituted a simplification of the task, keeping it more tractable. A further simplification is affected by keeping peak positions fixed during the fitting process, thus negating peak shifts due to conformational changes or the presence of numerous but weakly scattering small molecules. Even though there is a catch-all "other" category in recognition of the fact that Raman scattering does not all come from macromolecules, there is no guarantee that it would capture abundant small molecules, or less numerous strongly scattering ones, nor possible interactions between molecules. These effects might be too difficult to discern at present (see also below) and, due to the simplifications, they will be lumped into the most similar major groups.

The signal-to-noise ratio of measured spectra and inaccuracies of hyperspectral preprocessing might induce variations in peak amplitudes that can cause them to be erroneously grouped while inadequacies of hyperspectral preprocessing might cause interferents to be retained in spectra thus also hampering clustering. Smaller peaks might be especially vulnerable to these considerations. This may explain the separation of nucleic acid–related bands into Cluster 5 (mostly large peaks) and Cluster 1 (mostly small peaks), and the clustering heterogeneity observed for Cluster 3 (mostly the smallest peaks) as evident in Figure 4d. One might also deduce from this that using too many clusters will result in many clusters with small and perhaps unrelated peaks. In contrast, using too few clusters will result in the grouping together of many large peaks belonging to different macromolecules.

Furthermore, the experimental manipulations might be insufficient to induce differential changes in the sample components for effective clustering based on peak intensity trends. Conversely, simultaneous large changes (e.g., induced due to experimental manipulations or due to external conditions such as sample heterogeneity) could be present in all sample components such that the smaller changes that might permit discrimination between components are overwhelmed.

Another limitation derives from the current need to find suitable parameters for processing a specific data set. Some general approaches can be found outlined in the Methods section, as discussed here in Results and Discussion section or in previous work[25] while parameter values used here or previously[25] can serve to guide initial efforts. Though we are interested in developing fully automated algorithms,[9] this has not yet been achieved for the algorithms used here.

Regarding the utility of the method presented here, it is supplementary to two-dimensional cluster member spectroscopy[26] that also uses the clustering of trends as a separation procedure. However, it is complementary in the sense that a separate spectrum is generated from the fitted peaks within each cluster as opposed to two-dimensionally presenting the clustering of wavenumbers in a manner analogous to 2D-COS.[21] It is also supplementary to non-negative least squares and NNMF[13,22] by producing non-negative demixed results though using a very different approach. In the latter sense, it complements PCA[20] where the loadings of principal components can have both positive and negative features that make them difficult to interpret. Demixing with PCA is based on an analysis of the extent of dispersion of the spectral wavenumbers, whereas the current method is ultimately based on the similarity of dispersion of the wavenumbers arrived at through fitting. Finally, it is peak fitting that makes it complementary to all these methods by establishing peak parameters that can then be further analyzed individually, used for demixing through clustering, or used to create high-resolution spectra.[25]

## Conclusion

Rapid fitting of individual peaks in every spectrum of a Raman hyperspectral data set combined with *k*-means clustering of the peak amplitudes obtained from such fitting permits the use of peak parameters from all the peaks within a given cluster for the reconstruction of spectra representative of that cluster. These spectra are not distorted by unrelated overlapping peaks or noise, are less congested than those in the hyperspectral set, and improve peak identification and, potentially, individual macromolecule recognition. Such spectra might also produce spectra of cell components as they exist in the sample and the cluster centroids might contribute to an improved analysis of sample composition changes due to experimental effects. Though poor quality spectra and imperfections in spectral preprocessing and peak fitting can adversely affect the results, we have provided here a proof-of-concept of a novel mixture resolution method with different attributes than extant ones that thus can supplement them. In principle, these methods could be extended to the analysis of nonbiological complex materials with many proximal overlapping peaks.

### Acknowledgments

The authors thank Subin Mun for research assistance. The authors also acknowledge the Canadian Foundation for Innovation and the British Columbia Knowledge Development Fund for funding the purchase of instrumentation made available for this work through the UBC Laboratory for Molecular Biophysics. The authors also acknowledge the careful review of the manuscript by anonymous reviewers and the valuable comments that they provided.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors thank the Natural Sciences and Engineering Research

Council of Canada and the Canadian Institutes of Health Research for financial support.

## ORCID iDs

Shreyas Rangan https://orcid.org/0000-0001-8614-3894
Martha Z. Vardaki https://orcid.org/0000-0002-9624-8363
Michael W. Blades https://orcid.org/0000-0001-5272-4738
Robin F. B. Turner https://orcid.org/0000-0001-6786-7125

## References

1. A. Gutteridge, J.M. Rukstalis, D. Ziemek, M. Tié, et al. "Novel Pancreatic Endocrine Maturation Pathways Identified by Genomic Profiling and Causal Reasoning". PLoS One. 2013. 8 (2): e56024. 10.1371/journal.pone.0056024
2. T.C. Schulz, H.Y. Young, A.D. Agulnick, M.J. Babin, et al. "A Scalable System for Production of Functional Pancreatic Progenitors from Human Embryonic Stem Cells". PLoS One. 2012. 7(5): e37004. 10.1371/journal.pone.0037004
3. A. Rezania, J.E. Bruin, P. Arora, A. Rubin, et al. "Reversal of Diabetes with Insulin-Producing Cells Derived In Vitro from Human Pluripotent Stem Cells". Nat. Biotechnol. 2014. 32 (11): 1121–1133. 10.1038/nbt.3033
4. J.W. Chan, D.K. Lieu. "Label-Free Biochemical Characterization of Stem Cells Using Vibrational Spectroscopy". J. Biophotonics. 2009. 2(11): 656–668. 10.1002/jbio.200910041
5. S.O. Konorov, H.G. Schulze, B.K. Gage, T.J. Kieffer, et al. "Process Analytical Utility of Raman Microspectroscopy in the Directed Differentiation of Human Pancreatic Insulin-Positive Cells". Anal. Chem. 2015. 87(21): 10762–10769. 10.1021/acs.analchem.5b03295
6. S.O. Konorov, H.G. Schulze, N.J. Caron, J.M. Piret, et al. "Raman Microspectroscopic Evidence That Dry-Fixing Preserves the Temporal Pattern of Non-Specific Differentiation in Live Human Embryonic Stem Cells". J. Raman Spectrosc. 2011. 42 (4): 576–579. 10.1002/jrs.2769
7. H.G. Schulze, S.O. Konorov, N.J. Caron, J.M. Piret, et al. "Assessing Differentiation Status of Human Embryonic Stem Cells Noninvasively Using Raman Microspectroscopy". Anal. Chem. 2010. 82(12): 5020–5027. 10.1021/ac902697q
8. C. Krafft. "Raman Spectroscopy and Microscopy of Cells and Tissues". In: G.C.K. Roberts, editor. Encyclopedia of Biophysics. Berlin: Springer, 2013. Pp. 2178–2185. 10.1007/978-3-642-16712-6_121
9. H.G. Schulze, S. Rangan, J.M. Piret, M.W. Blades, R.F. Turner. "Developing Fully Automated Quality Control Methods for Preprocessing Raman Spectra of Biomedical and Biological Samples". Appl. Spectrosc. 2018. 72(9): 1322–1340. 10.1177/0003702818778031
10. H.G. Schulze, R.F. Turner. "Development and Integration of Block Operations for Data Invariant Automation of Digital Preprocessing and Analysis of Biological and Biomedical Raman Spectra". Appl. Spectrosc. 2015. 69(6): 643–664. 10.1366/14-07709
11. S.O. Konorov, H.G. Schulze, J.M. Piret, M.W. Blades, R.F.B. Turner. "Label-Free Determination of the Cell Cycle Phase in Human Embryonic Stem Cells by Raman Microspectroscopy". Anal. Chem. 2013. 85(19): 8996–9002. 10.1021/ac400310b
12. A. Pliss, A.N. Kuzmin, A.V. Kachynski, P.N. Prasad. "Nonlinear Optical Imaging and Raman Microspectrometry of the Cell Nucleus Throughout the Cell Cycle". Biophys. J. 2010. 99(10): 3483–3491. 10.1016/j.bpj.2010.06.069
13. K. Milligan, X. Deng, P. Shreeves, R. Ali-Adeeb, et al. "Raman Spectroscopy and Group and Basis-Restricted Non Negative Matrix Factorisation Identifies Radiation Induced Metabolic Changes in Human Cancer Cells". Sci. Rep. 2021. 11: 3853. 10.1038/s41598-021-83343-5
14. S. Rangan, S. Kamal, S.O. Konorov, H.G. Schulze, et al. "Types of Cell Death and Apoptotic Stages in Chinese Hamster Ovary Cells Distinguished by Raman Spectroscopy". Biotechnol. Bioeng. 2018. 115(2): 401–412. 10.1002/bit.26476
15. R. Luo, F. Wei, S. Huang, Y. Jiang, et al. "Real-Time, Label-Free Detection of Local Exocytosis Outside Pancreatic β Cells Using Laser Tweezers Raman Spectroscopy". Appl. Spectrosc. 2017. 71(3): 422–431. 10.1177/0003702816670911
16. N. Töpfer, M.M. Müller, M. Dahms, A. Ramoji, et al. "Raman Spectroscopy Reveals LPS-Induced Changes of Biomolecular Composition in Monocytic THP-1 Cells in a Label-Free Manner". Integr. Biol. 2019. 11(3): 87–98. 10.1093/intbio/zyz009
17. S.O. Konorov, H.G. Schulze, J.M. Piret, R.F.B. Turner, M.W. Blades. "Evidence of Marked Glycogen Variations in the Characteristic Raman Signatures of Human Embryonic Stem Cells". J. Raman Spectrosc. 2011. 42(5): 1135–1141. 10.1002/jrs.2829
18. H.G. Schulze, S.O. Konorov, J.M. Piret, M.W. Blades, R.F.B. Turner. "Empirical Factors Affecting the Quality of Non-Negative Matrix Factorization of Mammalian Cell Raman Spectra". Appl. Spectrosc. 2017. 71(12): 2681–2691. 10.1177/0003702817732117
19. A.N. Kuzmin, A. Pliss, A.V. Kachynski. "Biomolecular Component Analysis of Cultured Cell Nucleoli by Raman Microspectrometry: Biomolecular Analysis of Cultured Cell Nucleoli". J. Raman Spectrosc. 2013. 44(2): 198–204. 10.1002/jrs.4173
20. A. de Juan, J. Jaumot, R. Tauler. "Multivariate Curve Resolution (MCR). Solving the Mixture Analysis Problem". Anal. Methods. 2014. 6(14): 4964–4976. 10.1039/C4AY00571F
21. I. Noda. "Generalized Two-Dimensional Correlation Method Applicable to Infrared, Raman, and Other Types of Spectroscopy". Appl. Spectrosc. 1993. 47(9): 1329–1336. 10.1366/0003702934067694
22. P.-H. Chen, R. Shimada, S. Yabumoto, H. Okajima, et al. "Automatic and Objective Oral Cancer Diagnosis by Raman Spectroscopic Detection of Keratin with Multivariate Curve Resolution Analysis". Sci. Rep. 2016. 6(1): 20097. 10.1038/srep20097
23. Y. Xu, S. Deng, X. Li, Y. He. "A Sparse Unmixing Model Based on NMF and Its Application in Raman Image". Neurocomputing. 2016. 207: 120–130. 10.1016/j.neucom.2016.03.063
24. Y.-X. Wang, Y.-J. Zhang. "Nonnegative Matrix Factorization: A Comprehensive Review". IEEE Trans. Knowl. Data Eng. 2013. 25(6): 1336–1353. 10.1109/TKDE.2012.51
25. H.G. Schulze, S. Rangan, M.Z. Vardaki, M.W. Blades, et al. "Rapid Vector-Based Peak Fitting and Resolution Enhancement for Correlation Analyses of Raman Hyperspectra". Appl. Spectrosc. 2023. 77(8): 957–969. 10.1177/00037028231176805
26. H.G. Schulze, S. Rangan, M.Z. Vardaki, M.W. Blades, et al. "Two-Dimensional Clustering of Spectral Changes for the Interpretation of Raman Hyperspectra". Appl. Spectrosc. 2023. 77(8): 835–847. 10.1177/00037028221133851

27. J. MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations". Berkeley Symp. Math. Statist. Prob. 1967. 5.1(14): 281–297.
28. D. Steinley. "K-Means Clustering: A Half-Century Synthesis". Br. J. Math. Stat. Psychol. 2006. 59(1): 1–34. 10.1348/000711005X48266
29. R. Xu, D.C. Wunsch. "Clustering Algorithms in Biomedical Research: A Review". IEEE Rev. Biomed. Eng. 2010. 3: 120–154. 10.1109/RBME.2010.2083647
30. D.N. Stratis-Cullum, M.E. Farrell, E. Holthoff, D.L. Stokes, et al. "Spectroscopic Data in Biological and Biomedical Analysis". In: T. Vo-Dinh, editor. Biomedical Photonics Handbook. Boca Raton, Florida: CRC Press, 2014. Chapter 20, Pp. 610–839. 10.1201/b17290
31. A.C.S. Talari, Z. Movasaghi, S. Rehman, I. ur Rehman. "Raman Spectroscopy of Biological Tissues". Appl. Spectrosc. Rev. 2015. 50(1): 46–111. 10.1080/05704928.2014.923902
32. H.G. Schulze, R.B. Foist, K. Okuda, A. Ivanov, R.F.B. Turner. "A Small-Window Moving Average-Based Fully Automated Baseline Estimation Method for Raman Spectra". Appl. Spectrosc. 2012. 66(7): 757–764. 10.1366/11-06550
33. H.G. Schulze, R.F.B. Turner. "A Two-Dimensionally Coincident Second Difference Cosmic Ray Spike Removal Method for the Fully Automated Processing of Raman Spectra". Appl. Spectrosc. 2014. 68(2): 185–191. 10.1366/13-07216
34. H.G. Schulze, S. Rangan, J.M. Piret, M. Blades, R. Turner. "Smoothing Raman Spectra with Contiguous Single-Channel Fitting of Voigt Distributions: An Automated, High Quality Procedure". Appl. Spectrosc. 2019. 73(1): 47–58. 10.1177/0003702818794957
35. H.G. Schulze, S. Rangan, M.Z. Vardaki, M.W. Blades, et al. "Critical Evaluation of Spectral Resolution Enhancement Methods for Raman Hyperspectra". Appl. Spectrosc. 2022. 76 (1): 61–80. 10.1177/00037028211061174
36. J. De Gelder, K. De Gussem, P. Vandenabeele, L. Moens. "Reference Database of Raman Spectra of Biological Molecules". J. Raman Spectrosc. 2007. 38(9): 1133–1147. 10.1002/jrs.1734
37. A.D. Meade, C. Clarke, F. Draux, G.D. Sockalingum, et al. "Studies of Chemical Fixation Effects in Human Cell Lines Using Raman Microspectroscopy". Anal. Bioanal. Chem. 2010. 396(5): 1781–1791. 10.1007/s00216-009-3411-7
38. A.M. Piskorz, D. Ennis, G. Macintyre, T.E. Goranova, et al. "Methanol-Based Fixation Is Superior to Buffered Formalin for Next-Generation Sequencing of DNA From Clinical Cancer Samples". Ann. Oncol. 2016. 27(3): 532–539. 10.1093/annonc/mdv613
39. H.P.J. Bonarius, V. Hatzimanikatis, K.P.H. Meesters, C.D. De Gooijer, et al. "Metabolic Flux Analysis of Hybridoma Cells in Different Culture Media Using Mass Balances". Biotechnol. Bioeng. 1996. 50(3): 299–318. 10.1002/(SICI)1097-0290(19960505)50:3&lt;299::AID-BIT9>3.0.CO;2-B
40. F. Feijó Delgado, N. Cermak, V.C. Hecht, S. Son, et al. "Intracellular Water Exchange for Measuring the Dry Mass, Water Mass and Changes in Chemical Composition of Living Cells". PLoS One. 2013. 8(7): e67590. 10.1371/journal.pone.0067590
41. J.D. Jang, J.P. Barford. "Effect of Feed Rate on Growth Rate and Antibody Production in the Fed-Batch Culture of Murine Hybridoma Cells". Cytotechnology. 2000. 32(3): 229–242. 10.1023/A:1008169417980
42. J.M. Savinell, B.O. Palsson. "Network Analysis of Intermediary Metabolism Using Linear Optimization. I. Development of Mathematical Formalism". J. Theor. Biol. 1992. 154(4): 421–454. 10.1016/S0022-5193(05)80161-4
43. L. Xie, D.C. Wang. "Applications of Improved Stoichiometric Model in Medium Design and Fed-Batch Cultivation of Animal Cells in Bioreactor". Cytotechnology. 1994. 15(1–3): 17–29. 10.1007/BF00762376
44. C. Zupke, G. Stephanopoulos. "Intracellular Flux Analysis in Hybridomas Using Mass Balances and In Vitro 13C NMR". Biotechnol. Bioeng. 1995. 45(4): 292–303. 10.1002/bit.260450403