Received: 8 November 2018 | Revised: 8 February 2019 | Accepted: 12 February 2019

DOI: 10.1002/jbio.201800415

JOURNAL OF BIOPHOTONICS

FULL ARTICLE

An open-source code for Mie extinction extended multiplicative signal correction for infrared microscopy spectra of cells and tissues

Johanne H. Solheim[1]* | Evgeniy Gunko[1,2] | Dennis Petersen[3] | Frederik Großerüschkamp[3] | Klaus Gerwert[3] | Achim Kohler[1]

[1]Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway

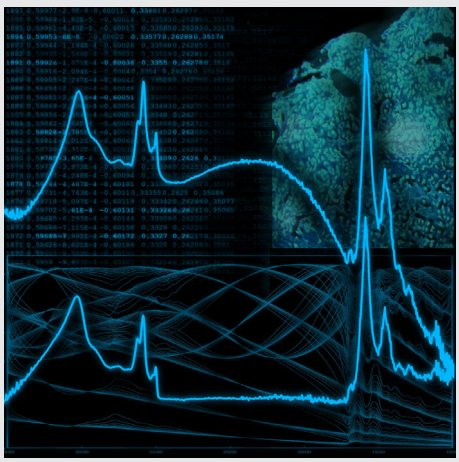
[2]Faculty of Radiophysics and Computer Technologies, Department of System Analysis and Computer Modeling, BY-Belarusian State University (BY-BSU), Minsk, Republic of Belarus

[3]Department of Biophysics and Protein Research Unit within Europe (PURE), Ruhr University Bochum, Bochum, Germany

***Correspondence**
Johanne H. Solheim, Faculty of Science and Technology, Norwegian University of Life Sciences, Drøbakveien 31, 1432 Ås, Norway.
Email: johanne.heitmann.solheim@nmbu.no

Funding information
Norges Miljø- og Biovitenskapelige Universitet

Infrared spectroscopy of single cells and tissue is affected by Mie scattering. During recent years, several methods have been proposed for retrieving pure absorbance spectra from such measurements, while currently no user-friendly version of the state-of-the-art algorithm is available. In this work, an open-source code for correcting highly scatter-distorted absorbance spectra of cells and tissues is presented, as well as several improvements of the latest version of the Mie correction algorithm based on extended multiplicative signal correction (EMSC) published by Konevskikh et al. In order to test the stability of the code, a set of apparent absorbance spectra was simulated. To this purpose, pure absorbance spectra based on a Matrigel spectrum are simulated. Scattering contributions where obtained by mimicking the scattering features observed in a set of experimentally obtained spectra . It can be concluded that the algorithm is not depending strongly on the reference spectrum used for initializing the algorithm and retrieves well the underlying pure absorbance spectrum. The calculation time of the algorithm is considerably improved with respect to the resonant Mie scattering EMSC algorithm used by the community today.



KEYWORDS
extended multiplicative signal correction, infrared microscopy, Mie scattering, open-source algorithm, preprocessing

1 | INTRODUCTION

Infrared spectroscopic imaging has proven successfully its capability to identify the chemical fingerprint of cells and tissues without having a degenerative effect on the sample [1–4]. In the last decade, application such as automatic identification of single cells and cancer cells in human tissue was established [1, 5–7]. In these works, it turned out that scattering effects originating from Mie scattering have a strong influence on the absorption spectra and data analysis [8]. Mie scattering occurs if spherical morphological structures are of comparable size as the incident radiation.

For infrared spectroscopy and biological samples, the incident radiation is of the size of single cells and cell nuclei. Thereby, the amount of intensity scattered away depends strongly on the wavelength of the electromagnetic radiation, the size parameters and refractive index of the scatterer. The scattered radiation that does not reach the detector leads to apparent absorption signatures in the measured absorbance spectrum. The preprocessing of highly distorted spectra is challenging because Mie scattering and absorption are highly entangled. In general, scattering and absorption phenomena are difficult to separate, as the two are mutual dependent on each other.

During the recent years, several algorithms have been developed and published that address the separation of scattering and absorption phenomena in apparent absorbance spectra of single cells [8–17]. Kohler et al developed an extended multiplicative signal correction (EMSC)-based algorithm for Mie scatter baseline correction [10], which was extended by Bassan et al. for resonant Mie scattering [8]. In the algorithm by Bassan et al., a wavelength-dependent real part of the refractive index was added to the approximation formula by van de Hulst for nonabsorbing spheres [18]. It has been observed that this algorithm had the tendency to produce corrected spectra which are contaminated by chemical features of the reference spectra. The contamination decreases as the number of iterations is increased, which leads to a compromise between speed and accuracy. Further improvements of the Mie correction algorithm was achieved by replacing the approximation formula for nonabsorbing spheres by the approximation of van de Hulst for spheres with complex refractive index, which is described in our recent publications [19, 20]. This approximation formula calculates the complete extinction, rather than only the scattering, and we therefore suggest to call the new algorithm Mie extinction EMSC (ME-EMSC). This algorithm corrects the broad Mie scatter oscillations in the apparent absorbance spectra of single cells and dispersive effects that are due to absorbance resonances.

In the paper at hand, we present and publish an open-source and user-friendly MATLAB code for retrieving pure absorbance spectra from highly distorted infrared absorbance spectra of cells and tissues. The code can be downloaded from GitLab,[a] and the paper provides clear directions for the use of the code. Further optimization of the algorithm published by Konevskikh et al. [19, 20] has been achieved by a number of improvements related to stability and speed. The algorithm is validated with a set of simulated apparent absorbance spectra, and we show that the algorithm converges quickly and toward a pure absorbance spectrum with the complete chemical information. Furthermore, we compare the performance of the implementation at hand and the latest version of the Mie scattering correction of Bassan et al. (RMieS-EMSC v3 and v5).

1. https://gitlab.com/BioSpecNorway/me-emsc

2 | THEORY

In this section, the ME-EMSC algorithm is described step by step. This is followed by a description of the method for simulating apparent absorbance spectra used for the validation of the algorithm. Details on Mie scattering in infrared spectroscopy of single cells and tissues, and the ME-EMSC model, can be found in Supporting information S1.

2.1 | The ME-EMSC algorithm

The underlying idea of the ME-EMSC algorithm is to retrieve the pure absorbance spectrum in an iterative process. The algorithm is initiated by selecting a reference spectrum, which has chemical features that are in general different from the underlying, and to be estimated, pure absorbance spectrum. The differences are within the chemical variability one expects for the data set. The reference spectrum is our first best guess for the underlying pure absorbance. The logic of the algorithm is such that the reference spectrum is updated after each iteration to a gradually better estimation of the true pure absorbance spectrum. The latest version of the algorithm, which is the basis for the code presented in this paper, was proposed by Konevskikh et al. [19]. The algorithm and further improvements are presented in the following. Figure 1 shows a schematic representation of the algorithm, where the red boxes mark improvements that are new with respect to the latest published version by Konevskikh et al. [20]. In the following, we explain each step of the algorithm shown in Figure 1, by referring explicitly to the boxes shown in the figure.

2.1.1 | Initialization

The algorithm is initiated by first selecting a reference spectrum. As reference spectrum, we may choose a standard reference spectrum, or a spectrum of the data set that is quasi scatter free. In imaging, for example, a major part of the spectra is scatter free and therefore they may be good candidates for reference spectra. As a standard spectrum for correcting spectra of cells and tissues, the Matrigel spectrum, either from Reference [8] or the one provided with this paper, can be used. The new Matrigel spectrum can be downloaded from the GitLab repository where the code is published, and the documentation can be found in Section 3 of the supporting material. In this paper, the Matrigel spectrum of Bassan et al. [8] will be used. This choice was made for comparing the new suggested algorithm with earlier versions of Mie correction algorithms which were based on the Matrigel spectrum. In order to align the reference spectrum such that the ranges used for the parameters a and h can be standardized, we decided to normalize the reference spectrum with respect to amide I. In addition, it is important that the reference spectrum is baseline corrected. Therefore, we suggest to select a baseline-free reference spectrum or to perform a baseline correction before the scaling with respect to

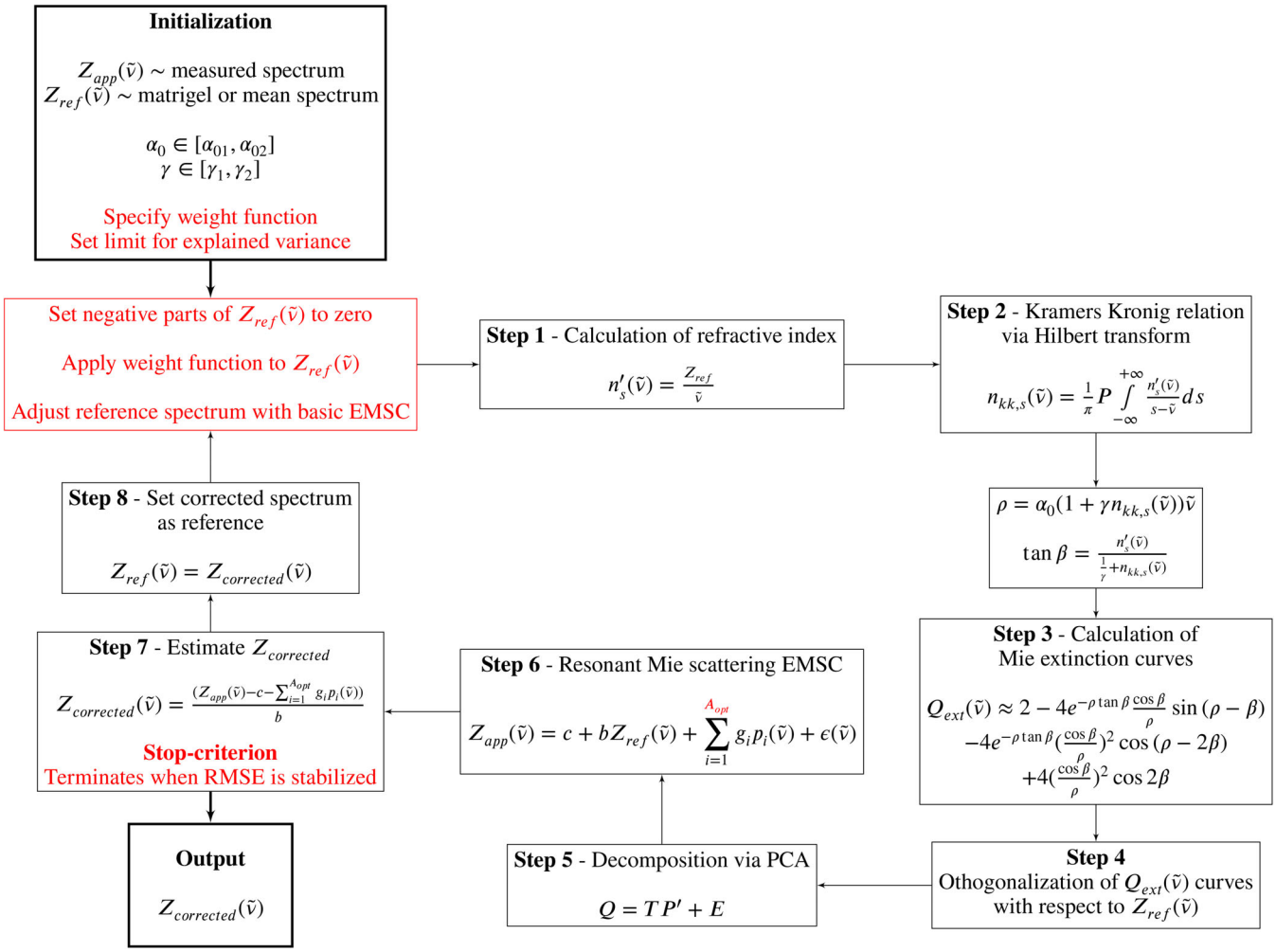


FIGURE 1 Schematic representation of the fast resonant Mie scatter correction algorithm. Red markings indicate changes done with respect to the algorithm of Konevskikh et al. [20]

the amide I. The scaling of the reference spectrum allows using the same parameter range even if multiple reference spectra are used. Ranges for α_0 and γ need to be specified once for a data set. They are set by defining a parameter range for n_0, a and h. The program sets then the parameter range for α_0 and γ automatically.

Step 1: Estimation of $\boldsymbol{n'_s}$

The reference spectrum is used for estimating the scaled imaginary part of the refractive index as

$$n'_s(\tilde{\nu}) = \frac{Z_{\text{ref}}}{\tilde{\nu}} \tag{1}$$

which is related to the imaginary part n' by

$$n'_s = f \cdot n' \qquad \text{where} \quad f = \frac{\ln(10)}{4\pi d_{eff}} \tag{2}$$

Step 2: Estimation of $\boldsymbol{n_{kk}}$ by the Kramers-Kronig relation

The scaled fluctuating part of the real refractive index is estimated from the scaled imaginary part according to the Kramers-Kronig relation

$$n_{kk}(\tilde{\nu}) = \frac{2}{\pi} P \int_0^{+\infty} \frac{s \cdot n'(s)}{s^2 - \tilde{\nu}^2} \, ds \tag{3}$$

It can be shown that by considering the symmetry of the refractive index, the Kramers-Kronig relation is equivalent to the Hilbert transform [19], written as

$$n_{kk,s}(\tilde{\nu}) = \frac{1}{\pi} P \int_{-\infty}^{+\infty} \frac{n'_s(s)}{s - \tilde{\nu}} ds = -\frac{1}{\pi\tilde{\nu}} * n'_s(\tilde{\nu}) \tag{4}$$

where * denotes convolution. The Hilbert transform can be calculated via the fast Fourier transform, leading to a decrease in computational time by a factor of 100 compared to calculating the Kramers-Kronig integral [19].

Step 3: Calculation of the set of Mie extinction curves

With an estimation of n_s' and $n_{kk,s}$ at hand, Mie extinction curves can be calculated according to the van de Hulst approximation [18] given by

$$Q_{\text{ext}}(\tilde{\nu}) \approx 2 - 4e^{-\rho \tan\beta} \frac{\cos\beta}{\rho} \sin(\rho - \beta)$$

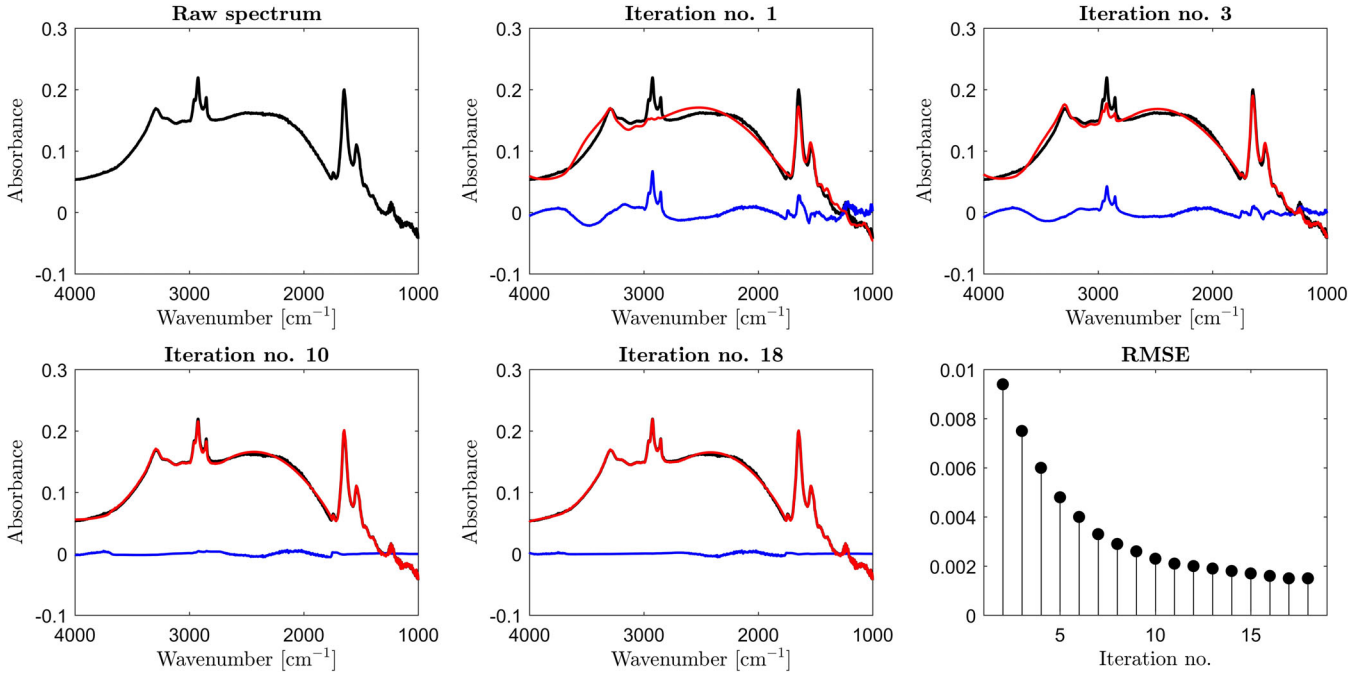


FIGURE 2 The ME-EMSC forward model. Correction of a measured lung cancer cell spectrum (in black) [10]. After each iteration, a better estimate for the pure absorbance is obtained, leading to a more precise prediction of the apparent absorbance spectrum (in red). The residuals are shown in blue. The RMSE decreases in each iteration (lower right)

$$-4e^{-\rho\tan\beta}\left(\frac{\cos\beta}{\rho}\right)^2\cos(\rho-2\beta)$$
$$+4\left(\frac{\cos\beta}{\rho}\right)^2\cos 2\beta \quad (5)$$

Dividing the parameter ranges for α_0 and γ into 10 distinct values, a total of 100 different extinction curves Q_{ext} are obtained.

Step 4: Orthogonalization of $\boldsymbol{Q}_{\text{ext}}$ with respect to Z_{ref}

As the extinction curves Q_{ext} are approximately proportional to the apparent absorbance spectrum Z_{app}, both Q_{ext} and Z_{app} contain attenuation features due to both chemical absorption and scattering. Chemical information in the form of the imaginary part of the refractive index n_s' is entering Q_{ext} as Z_{ref}. One of the key features of the EMSC model is that each spectrum is modeled around the reference spectrum. The reference spectrum itself in the EMSC model allows to estimate the scaling parameter b in the model given in Equation 7, which refers to the effective optical path length. In order to avoid competition between the parameter estimation for the reference spectrum and other model parameters, the 100 different extinction curves Q_{ext} are orthogonalized with respect to the reference spectrum Z_{ref}. This is equivalent to regressing the extinction curves on the reference spectrum Z_{ref} and removing the estimated contribution from the reference spectrum Z_{ref} to the extinction curves.

Step 5: Meta-modeling of $\boldsymbol{Q}_{\text{ext}}$ for a parameter range via principal component analysis (PCA)

For each parameter pair of α_0 and γ, the extinction efficiency Q_{ext} is calculated according to the Van De Hulst formula in Equation 5 as a function of the wavenumber $\tilde{\nu}$. All obtained extinction efficiencies are collected in the matrix $\boldsymbol{Q}$ and are decomposed via PCA according to

$$\boldsymbol{Q} = \boldsymbol{T_A}\boldsymbol{P_A}' + \boldsymbol{E_A} \quad (6)$$

where $\boldsymbol{T_A}$ and $\boldsymbol{P_A}'$ are the scores and loadings, respectively, and $\boldsymbol{E_A}$ expresses the residual. The parameter A refers to the number of components included. The loadings in $\boldsymbol{P_A}'$ are representing a meta-model of Q_{ext}, and are included as model spectra in the EMSC model.

Step 6: ME-EMSC

The first A loadings from the PCA model are included in the ME-EMSC model as model functions p_i. The complete model is given by

$$Z_{app}(\tilde{\nu}) = c + bZ_{\text{ref}}(\tilde{\nu}) + \sum_{i=1}^{A_{opt}} g_i p_i(\tilde{\nu}) + \epsilon(\tilde{\nu}) \quad (7)$$

The parameters c, b and g_i are estimated by least squares regression. The residual ϵ contains the unmodelled part of the apparent absorbance spectrum. Thus, ϵ is expected to contain chemical differences between the reference spectrum and the true pure absorbance spectrum, unmodelled noise and unmodelled scatter contribution. For more details on the ME-EMSC model, see supporting information.

Step 7: Estimate $\boldsymbol{Z}_{\text{corrected}}$

After estimation of the parameters c, b and g_i by least squares regression, a corrected spectrum is obtained according to

$$Z_{\text{corrected}}(\tilde{\nu}) = \frac{Z_{\text{app}}(\tilde{\nu}) - c - \sum_{i=1}^{A_{\text{opt}}} g_i p_i(\tilde{\nu})}{b} \quad (8)$$

Step 8: Update reference spectrum

Calculating the corrected spectrum according to Equation 8 is equivalent to updating the reference spectrum by adding the scaled residuals of Equation 7 according to

$$Z_{\text{corrected}}(\tilde{\nu}) = Z_{\text{ref}} + \frac{\epsilon(\tilde{\nu})}{b}. \quad (9)$$

The scaled residuals represent the unmodelled part of Equation 7. Assuming that this unmodelled part is mainly due to chemical differences between the reference spectrum and the true underlying pure absorbance spectrum, the next estimate of the corrected spectrum in Equation 9 is expected to be closer to the true pure absorbance spectrum.

For the next iteration, the new estimate of the corrected spectrum serves as the updated reference spectrum. Thus, the updated reference spectrum is expected to be slightly closer to the true pure absorbance spectrum. In addition, for the next iteration step, both the negative parts of the reference spectrum Z_{ref} and the negative parts of the imaginary part of the refractive index n' are set to zero, as negative parts of the imaginary part of the refractive index and the pure absorbance spectrum refer to a nonphysical situation.

2.1.2 | Stop criterion

In general, we expect the residual to decrease after each iteration step. This is due to the fact that every update of the reference spectrum brings the reference spectrum closer to the true pure absorbance spectrum underlying the measured apparent absorbance.

With a better estimate for the pure absorbance, the predicted apparent absorbance spectrum, given by

$$Z_{\text{predicted}} = c + bZ_{\text{ref}}(\tilde{\nu}) + \sum_{i=1}^{A_{\text{opt}}} g_i p_i(\tilde{\nu})$$

approaches the measured apparent absorbance spectrum, while the residual decrease with the number of iterations (Figure 2).

Konevskikh et al. [20] propose to terminate the iterative process when the root mean square error (RMSE) has reached the threshold of RMSE<10^{-4}. The RMSE is calculated according to

$$\text{RMSE} = \sqrt{\frac{1}{N}\left(Z_{\text{ref}} - Z_{\text{predicted}}\right)^2} \quad (10)$$

where N is the number of wavenumber channels.

2.2 | Simulation of data sets

In order to test the stability of the resonant Mie algorithm, a set of apparent absorbance spectra was simulated. The simulated spectra were subject to the following two requirements:

1. The underlying pure absorbance spectra need to be known for validating the pure absorbance spectra that were retrieved from the simulated apparent absorbance spectra.
2. The scattering features of the simulated apparent absorbance spectra were required to resemble scattering features observed in experimentally obtained spectra.

The motivation for the last requirement was to take into account that measured absorbance spectra are obtained from samples that are not perfect and homogeneous spheres. Therefore, the simulation of an apparent absorbance spectrum that is simply based on a perfect sphere usually shows features that are very different from measured apparent absorbance spectra of cells. We decided to base the simulation of scatter features on experimentally obtained spectra. The experimentally obtained spectra that were used for estimating scatter contributions included 59 infrared spectra obtained from lung cancer cells. The experimental setup used to acquire the spectra can be found in Konevskikh et al. [10]. Apparent absorbance spectra were simulated according to:

$$Z_{\text{app}}(\tilde{\nu}) = c + bZ_{\text{pure}}(\tilde{\nu}) + \sum_{i=1}^{A} g_i p_i(\tilde{\nu}) \quad (11)$$

where Z_{pure} is a simulated pure absorbance spectrum. It is important to note that Equation 11 does not represent an additive model where the pure absorbance spectrum Z_{pure} is simply added to scatter baselines described by the third term of the right-hand side of Equation 11. This term is obtained by simulating a set of scatter extinctions using the real and imaginary parts of the refractive index calculated from the pure absorbance spectrum Z_{pure}. The obtained set of scatter extinctions is then orthogonalized with respect to Z_{pure}. The details of the establishment of the simulated spectra are explained in the following.

In order to base the physical attenuation on measured spectra, scatter contributions and chemical contributions in measured spectra were first estimated employing the resonant Mie scatter algorithm. The parameters c, b and g_i were taken from the last iteration in the correction, and used as input for Equation 11. For the simulation of pure absorbance spectra, we used the Matrigel spectrum [9] as a template. In order to obtain a set of pure absorbance spectra with chemical variation, the Matrigel spectrum was decomposed into a set of Lorentz lines [19]. Band positions were then systematically and randomly shifted either to the left or to the right (± 1 cm^{-1}), and peak heights were changed by $\pm 20\%$. This resulted in a set of pure absorbance spectra, which resemble the Matrigel spectrum, but with slightly different chemical information. In Figure 3A, a simulated pure absorbance spectrum is shown in red together with the Matrigel spectrum in black. Differences in band heights can be seen all over the spectrum, while they are especially strong in the region between 1500.0 and 1000.0 cm^{-1}. Two sets of pure absorbance spectra were simulated, representing two

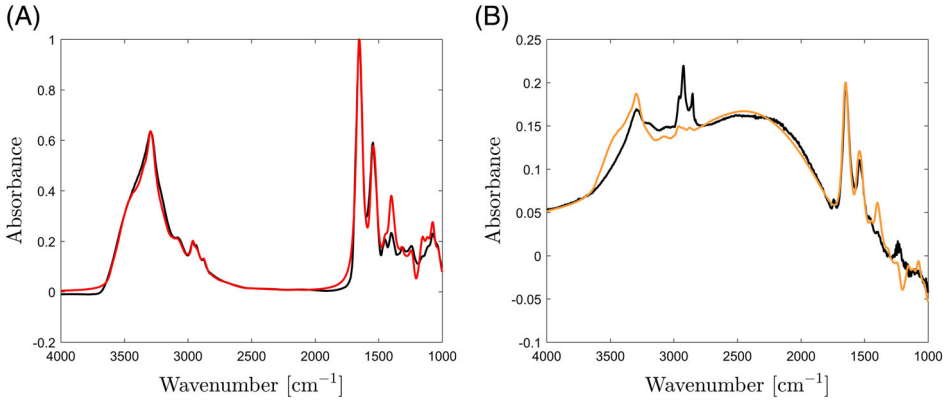


FIGURE 3 Simulation of pure and apparent absorbance spectra. A, Simulated pure absorbance spectrum in red, and the Matrigel spectrum in black. B, Simulated apparent absorbance spectrum in orange is shown together with the measured spectrum from which the scatter parameters are obtained for the simulation

chemically different groups, A and B, with some random variation within each group. Group A and B consisted of 25 spectra each.

The scatter effects described by the third term of the right-hand side of Equation 11 are obtained by first calculating n' and n_{kk} using the simulated pure absorbance spectra Z_{pure} as input for Equation 6 in the supporting information. Then, a set of Q_{ext} was simulated for a parameter range for a, n_0 and h. Here the same parameter ranges were used as for estimating the parameters c, b and g_i from the measured spectra. The obtained set of Q_{ext} was then decomposed by PCA. The loadings from the PCA, together with the parameters c, b and g_i obtained from the measured spectrum, were then used as input into Equation 11.

In order to make sure that the simulated apparent absorbance spectra resemble measured spectra, Z_{pure} was scaled appropriately. Scaling of Z_{pure} was done such that the features of the chemical absorbance were as strongly expressed as they were in a typical measured spectrum.

Simulations according to this procedure resulted in apparent absorbance spectra with scatter contributions that closely resembled the scatter contributions of experimentally obtained spectra, while the underlying pure absorbance spectrum is known a priori. An example is shown in Figure 3B. The simulated apparent absorbance spectrum is shown in orange, while the experimentally obtained apparent absorbance spectrum is shown in black. Each of the 59 infrared-spectra obtained from lung cancer cells [10] were used as a template for estimating scatter contributions. The scatter contributions where used as input for simulating apparent absorbance spectra together with chemical contribution from one pure absorbance from groups A and B, respectively. This resulted in a total of 118 simulated spectra. In the figures of this paper, the pure absorbance spectrum from group A is always plotted in dark red, while the pure absorbance spectrum from group B is always plotted in dark blue. Simulated apparent absorbance spectra, and the corresponding corrected spectra, are plotted in orange and light blue, respectively. The Matrigel spectrum is plotted in black.

3 | RESULTS AND DISCUSSION

In the paper at hand, the algorithm of Konevskikh et al. [20] has been further improved. The changes relate to the optimization and stabilization of the algorithm and include

- Optimization of the number of principal components used in the model
- Optimization of the stop criterion
- Weighting of the reference spectrum
- Guaranteed positivity of the reference spectrum
- Scaling the reference spectrum in each iteration

The algorithm is validated by the use of a simulated data set. Furthermore, the dependency of the retrieved pure absorbance spectrum on the reference spectrum used, and the retrieval of the true amide I peak position are reviewed. Finally, the sensitivity toward initialization parameters is discussed. In order to give guidance to the user of the provided algorithm, we explain every single step of usage of the algorithm by an example.

3.1 | Proposed changes to the algorithm

3.1.1 | Setting the optimal number of principal components

As described in the previous section, a small number A of PCA loadings are used in the RMieS-EMSC model in order to compress the set of extinction curves Q_{ext}. The number of loadings A has an impact on the precision of the model, its stability and the computational time. Thus, A should be chosen carefully and such that the Mie oscillations are represented precisely.

To achieve this, the level of explained variance represented by the number of loadings A is set when initiating the algorithm. According to our experience with different datasets, a good level of explained variance is between 99.96% and 99.99%. The optimal level of the explained variance may differ when a different grid is set for α_0 and γ. How the level of the explained variance and the grid resolution are

related in detail will be studied further elsewhere. The number of loadings A is set in the first iteration, and is calculated based on the level of the explained variance. Typically, A is estimated to assume numbers from $A = 7$ to $A = 9$. If A is too low, the Mie oscillations will not be represented precisely, resulting in an unstable baseline correction, which still contains scattering features. In this case, the level of explained variance has to be increased. When the number of components reaches a certain level, a further increase results in negligible contributions to the model, and the correction does not change. However, increasing the number of loadings to a very high level, this may result in noisy loadings and an instability related to the noise level.

3.1.2 | Revised stop criterion

After each iteration of the Mie correction algorithm, it is expected that the RMSE of the prediction decreases. Nevertheless, the algorithm does not necessarily converge to zero or a very low RMSE: If the apparent absorbance spectrum contains features that are not described by the EMSC model used, and if these features are strong, then the RMSE may assume a relatively high value. In the previous version of the algorithm, an absolute RMSE limit of 10^{-4} was used. This may lead to a high number of iterations without achieving a reasonable correction. We propose therefore a stop criterion that is based on the convergence of the RMSE. We terminate the algorithm when the RMSE does not change substantially. The algorithm is also terminated if the RMSE starts to increase. In addition, we allow setting a maximum RMSE that can be used to determine if a correction was successfully completed or not. The maximum RMSE can be set by the user. As default it is set to infinity.

An option for overruling the stop criterion is implemented in the published code. This is done by setting a fixed number of iterations. Caution should be taken when applying this option, as with only one or a few iterations, a corrected pure absorbance spectrum is not achieved. The algorithm relies on the iterative process of updating the reference spectrum, and gradually retrieving the underlying pure absorbance. After only a few iterations, the corrected spectrum will resemble the reference spectrum significantly.

3.1.3 | Weighting of the reference spectrum

Absorbance signals from the chemically inactive regions of the sample, for example, from carbon dioxide molecules in the air at around ~2300 cm^{-1}, should not be modeled as properties of the sample. These absorption modes are not deformed by Mie scattering. Therefore, we suggest to downweight the chemically inactive regions by a weight function. This is not equivalent to standard weighting in EMSC, or weighted least squares, as the weighting is applied only to the reference spectrum. The weight function can be seen in Figure 9C and D, and details on how the weight function is implemented can be found in the supporting information. The result of employing a weight function is a flat baseline with less disturbances.

3.1.4 | Guaranteed positive reference spectrum

Konevskikh et al. [20] suggested to set negative parts of the imaginary refractive index to zero due to physical considerations. For the same reason, we suggest that negative parts of the reference spectrum used in the EMSC are set to zero, such that a more physical reference spectrum is used in the modeling. This means in practice that we set the negative parts of the reference spectrum to zero, from which thereafter the imaginary part of the refractive index is calculated according Equation 1, ensuring thereby already positivity of the imaginary part of the refractive index. The imaginary part of the refractive index does then not contain any negative parts since it is directly calculated from the reference spectrum in each iteration step according to Equation 6 in the supporting material.

3.1.5 | Scaling the reference spectrum

Konevkikh et al. [20] proposed to normalize the Matrigel spectrum when initializing the algorithm. In order to ensure that the scaling of the reference spectrum in each iteration is maintained, we suggest applying a basic EMSC on the reference spectrum with respect to the initial reference spectrum in each iteration step. When the reference spectrum is not scaled in each iteration step, small scaling differences may occur. The scaling of the reference spectrum in each iteration step is done by dividing Z_{ref} by b from the basic EMSC according to

$$Z^k_{\text{ref}}(\tilde{\nu}) = c + bZ^0_{\text{ref}}(\tilde{\nu}) + d\tilde{\nu} + e\tilde{\nu}^2 + \epsilon(\tilde{\nu}) \qquad (12)$$

where Z^k_{ref} is the reference spectrum in iteration no. k, and Z^0_{ref} is the reference spectrum used for initializing the algorithm.

The scaling of the reference spectrum in each iteration step is to avoid a small but gradual scaling deviations in the reference spectrum in each iteration step. It is observed that the adjustments are very small in each iteration, which is taken as an indication of stability of the algorithm.

3.2 | Validation of the algorithm

3.2.1 | Simulation of pure absorbance spectra

In order to be able to determine whether the spectra can be correctly classified after correction, two sets of pure absorbance spectra were simulated, using the method described in Section 2.2. In Figure 4A), the simulated pure absorbance spectra are shown. In total, 50 pure absorbance spectra were simulated. The set of pure absorbance spectra is forming two chemically different groups as visual inspection of Figure 4A) shows. The spectra of the two different groups are plotted in red and blue, respectively. The Matrigel is plotted in black. PCA was used to investigate the simulated

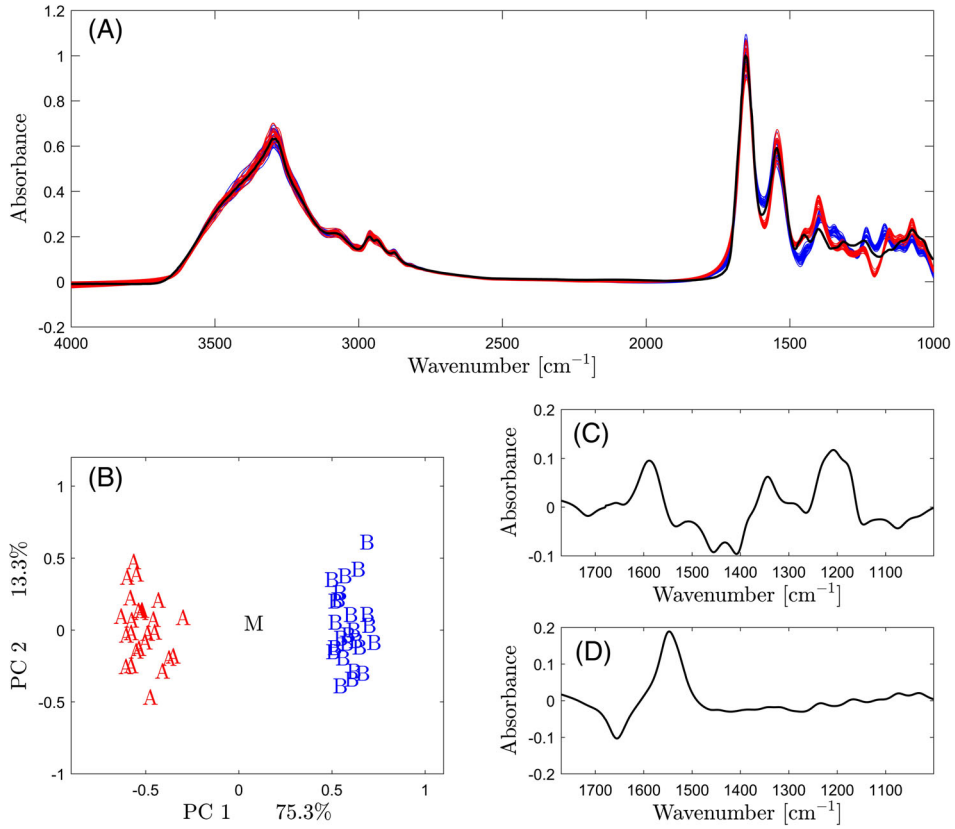


FIGURE 4 A, Simulated pure absorbance spectra, representing two chemically different groups. Group A is plotted in red, and group B in blue, and the Matrigel spectrum is plotted in black. B, Score plot of the first and second components of the PCA on the simulated pure absorbance spectra. Two clusters are obtained. This PCA will serve as a reference for evaluating the correction. (C and D) First and second loadings from the PCA, respectively. They show simulated chemical features

data. In Figure 4B), the score plot of the first and second components is shown for the PCA analysis on the simulated data set of pure absorbance spectra, including the Matrigel spectrum. We observe that the simulated data set shows a clear separation of the two chemically different groups, groups A and B, with some random variation within each group. The Matrigel spectrum is located between the two chemically different groups. Here and in the following, the PCA analysis is performed on the spectral region from 1000 and 1770 cm^{-1}. It is evident from Figure 4B) that the chemical differences between the groups result in two distinct clusters, both separated from the Matrigel spectrum, which is located in the middle. In Figure 4C and D, the first and second loading are shown. They show simulated chemical features. In the following, the score plot in Figure 4B) will be used as a reference to determine whether the correction of the simulated apparent absorbance spectra can be considered successful or not. In the following, we will only use one pure absorbance spectrum from each group and introduce different scatter effects that are obtained from a set of measured apparent absorbance spectra.

3.2.2 | Simulation of apparent absorbance spectra

In order to evaluate the algorithms ability to retrieve pure absorbance spectra, one spectrum from each group of simulated pure absorbance spectra was chosen and used as a pure absorbance spectrum for the simulation of apparent absorbance spectra with different scatter features. For the simulation of the apparent absorbance spectra, a variety of scatter features from measured spectra were used as described in Section 2.2. In total, 59 apparent absorbance spectra with different scattering contributions were simulated based on each pure absorbance spectrum. The simulated apparent absorbance spectra are shown in Figure 5A). Apparent absorbance spectra that were simulated with the pure absorbance spectrum from group A are plotted in orange, while the apparent absorbance spectra that were simulated with the pure absorbance spectrum from group B are plotted in light blue. A variety of scattering features can be observed. While all orange and light blue apparent absorbance spectra are based on one pure absorbance spectrum from groups A and B, respectively, each simulated apparent absorbance spectrum is based on the scatter features of one of 59 measured apparent absorbance spectra.

Figure 5B) shows the score plot of the first two principal components of a PCA performed on the set of apparent absorbance spectra. It can be seen that the first two components do not allow to group the apparent absorbance spectra according to the groups A and B. The samples are spread out in such a way that it is impossible to distinguish between

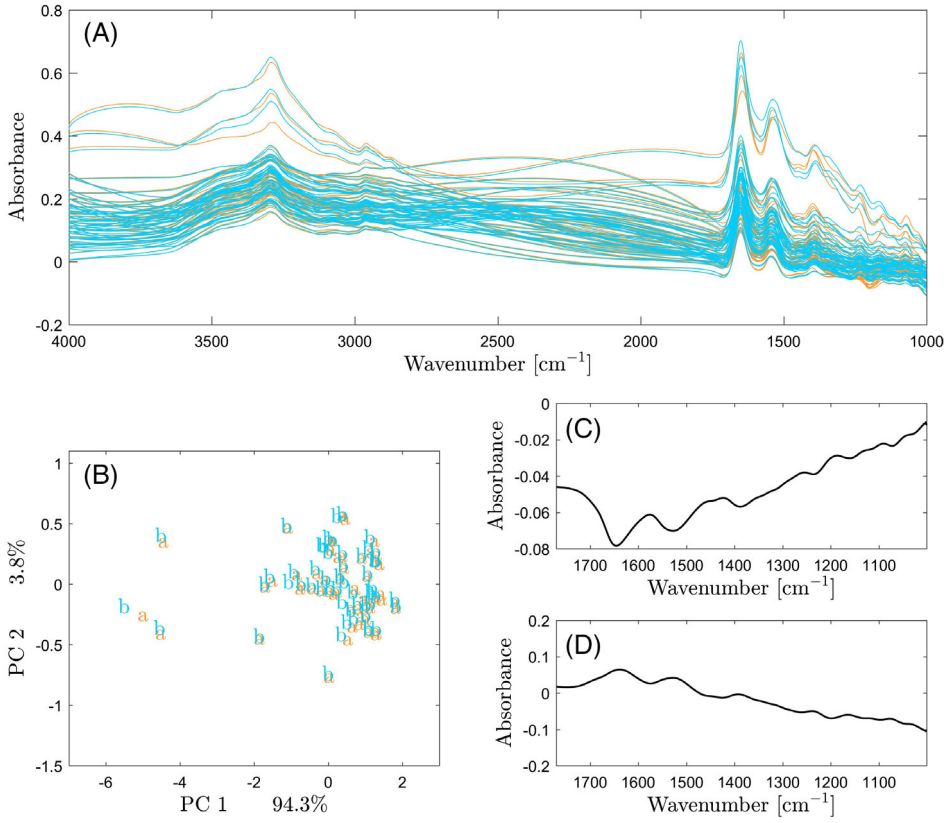


FIGURE 5 A, Simulated apparent absorbance spectra. Spectra with an underlying pure absorbance spectrum from group A are shown in orange, while for the light blue, a pure absorbance from group B is used. B, The score plot of the first and second components of the PCA performed on the apparent absorbance spectra shows that the clustering that was observed for the pure absorbance spectra cannot be achieved for the simulated apparent absorbance spectra. In C and D, the corresponding first and second loadings are shown. Both loadings reveal clear scatter features

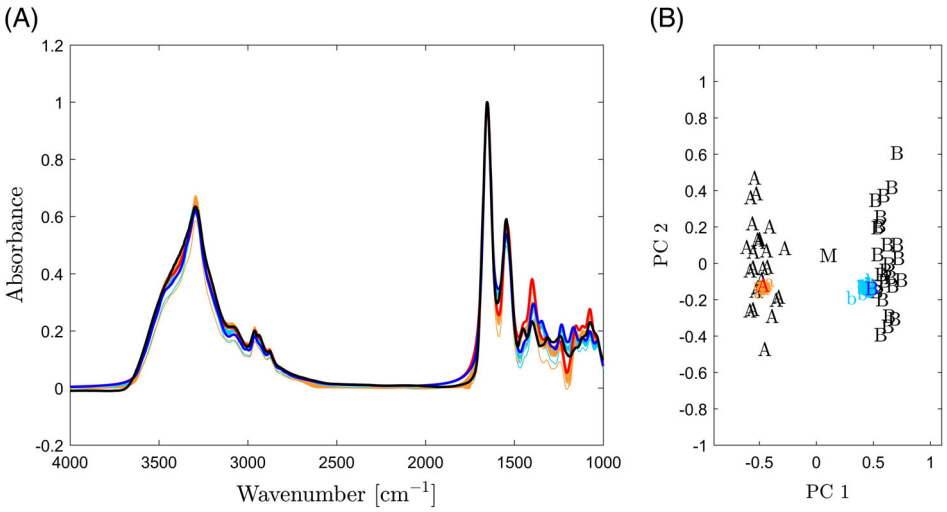


FIGURE 6 A, Corrected simulated apparent absorbance spectra in orange and light blue. They correspond to the pure absorbance spectra shown in red (from group A) and dark blue (from group B), respectively. The Matrigel spectrum is shown in black. B, Projection of the corrected apparent absorbance spectra into the score plot of the PCA of the pure absorbance spectra shown in Figure 4B. The corrected spectra cluster around the true pure absorbance spectra. Color coding corresponds to the color coding used in A. As a reference, the scores of all simulated pure absorbance spectra from group A and B are shown in black

the two groups. The corresponding first and second loading vectors of the PCA are shown in Figure 5C and D. It can be seen that the loadings describe scatter features rather than chemical features.

3.2.3 | Retrieval of pure absorbance spectra

The set of apparent absorbance spectra was corrected using the algorithm presented in this paper. The corrected spectra are shown in Figure 6A. The corrected spectra that are based

on the pure absorbance spectrum from group A are plotted in orange, and the corrected spectra that are based on the pure absorbance spectrum of group B in light blue. The pure absorbance spectrum of group A is plotted in red, while the pure absorbance spectrum of group B is plotted in dark blue. The Matrigel spectrum is plotted in black. It is obvious that the fast Mie scatter correction algorithm presented in this paper is able to restore the chemical features of the underlying pure absorbance spectra instead of resulting in corrected spectra that were contaminated by chemical features of the Matrigel spectrum. In order to evaluate the quality of the correction, the corrected spectra were projected into the score plot of the PCA of the simulated pure absorbance spectra of Figure 4B. The result is shown in Figure 6B. It can be observed that for both group A and group B, the corrected spectra cluster tightly around the pure absorbance spectrum that was used for simulations of apparent spectra with different scatter features. As mentioned previously, our earlier versions of the Mie correction algorithm suffered from being strongly affected by the reference spectrum. This can be clearly seen in Reference [9] in fig. 1, where the corrected spectra adapt features from the reference spectrum. From Figure 6B, it is evident that the algorithm presented in this paper does not suffer from this problem; the corrected spectra cluster around the true pure absorbance spectrum with a spread that is much lower than the spread within the two chemical groups A and B.

By means of classification with PCA, the algorithm has shown to be reliable, and the true pure absorbance spectrum is retrieved.

3.3 | Dependency on the reference spectrum

Earlier versions of the resonant Mie scattering algorithm revealed challenges related to the correction being strongly biased by the initial reference spectrum. It appeared that the corrected spectrum was dominated by features of the reference spectrum. This is evident in fig. 1 of Reference [9], where we can see that to the left of the O—H stretching peak at approximately 3400 cm^{-1}, the corrected spectrum has a shoulder which appeared in the reference spectrum but not in the measured raw spectrum. In addition, at around approximately 2900 cm^{-1}, corresponding to a lipid absorption band, the absorbance is much lower in the corrected spectrum than what is expected when inspecting visually the measured apparent absorbance spectrum. In the following, we will demonstrate that the algorithm presented in this paper, which is based on the algorithm proposed by Konevskikh et al. [19], does not suffer from the problem to be biased to a strong degree by the reference spectrum.

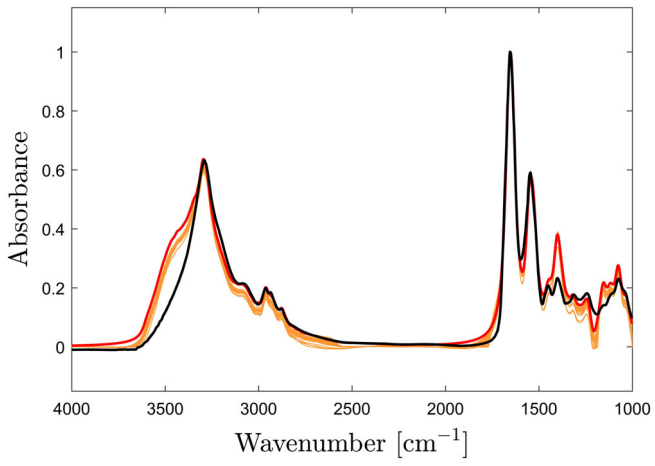


FIGURE 7 Correcting simulated apparent absorbance spectra with a modified version of the Matrigel spectrum (in black) as initial reference spectrum. Corrected spectra are shown in orange, and their underlying pure absorbance spectrum is shown in red

3.3.1 | Reference spectrum with altered O—H stretching region

In order to test the influence of the reference spectrum on the corrected spectra, the absorbance of the Matrigel spectrum was modified in the O—H stretching region. In Figure 7, the modified Matrigel spectrum is shown in black. Its absorption is lowered to the left of the peak at ca. 3300 cm^{-1}. The pure absorbance spectrum that was used for simulating apparent absorbance spectra is shown in red. The pure absorbance spectrum is based on the Matrigel spectrum, but contains chemical differences that were simulated as described in Section 2.2. We observe that the Matrigel spectrum in addition differs in the broad O—H stretching absorbance. This difference was introduced in order to investigate if the corrected spectrum adapts to the modified Matrigel spectrum in the O—H stretching region or not, a feature that was previously observed for the algorithm developed by Bassan et al. [9] (See, eg, fig. 1 in Reference [9]). When correcting the simulated apparent absorbance spectra with the black reference spectrum, the corrected pure absorbance spectra plotted in orange in Figure 7 was obtained. We observe that the true features of the pure absorbance spectrum were retrieved.

3.4 | Ability to retrieve the true amide I peak position

When introducing the resonant Mie model, Bassan et al. [9] showed that a more reliable peak position of the amide I absorption band at 1655 cm^{-1} could be retrieved. A shift in the amide I peak position occurs due to the so-called "dispersive effect," that is, the effect of a fluctuating real refractive index, caused by the frequency-dependent absorption. For the amide I absorption band, this results in a shift toward lower wavenumbers. This can be incorrectly interpreted as changes in the secondary structure of local segments in proteins. As the amide I peak position can play a major role in classification of cells and tissues, it is desired to retrieve the true peak position when preprocessing spectra. In this paper, we have demonstrated that the correction is less dependent on the reference spectrum. In order to demonstrate that the

retrieval of the true amide I peak position is a feature of the model, and not related to the reference spectrum, we investigate this matter closer. We start by correcting simulated apparent absorbance spectra, where a significant shift in amide I is present. Subsequently, simulated apparent absorbance spectra are corrected with a modified version of the Matrigel spectrum, where the amide I peak position is dislocated.

3.4.1 | Correcting simulated spectra with a significant shift in amide I

The lung cancer cell data set used for simulations does not show strong shifts in amide I peak position in the raw spectra. When correcting the amide I position in the measured spectra, the peak position was moved from on average 1649.8 ± 2.1 to 1651.8 ± 1.7 cm^{-1}. Due to the small displacement of the absorption band, the shifts in the simulated apparent absorbance spectra were also rather small. During the simulations, the amide I position was moved from 1653.0 cm^{-1} in the pure absorbance spectra, to on average 1649.5 ± 1.6 cm^{-1} in the apparent absorbance spectra. The correction brought the peak back to 1652.7 ± 0.3 cm^{-1}.

In order to demonstrate this for cases where the shift in the peak position is greater, new simulations were done based on measured data of breast cancer cells provided by Nick Stone. A description of the data set can be found in the supporting material. In this data set, the amide I peak position is found at 1635.3 ± 4.8 cm^{-1}, and corrections moved it to 1651.0 ± 1.2 cm^{-1}. Figure 8 shows an example of a simulated spectrum, where the amide I position is shifted from 1653.0 cm^{-1} in the pure absorbance spectrum, to 1643.0 ± 1.9 cm^{-1} in the apparent absorbance spectrum. When correcting the apparent absorbance spectra, all amide I peak positions were shifted back to 1650.8 cm^{-1}. As the correction is shown to be less dependent on the reference spectrum, this is taken as a strong indicator that the retrieval of the true amide I peak position is related to features of the model rather than the reference spectrum. In the supporting material, it is further shown that even in cases where the amide I peak position is shifted in the reference spectrum, the correction does not adapt to this shift. It is concluded that the retrieval of the true peak position is a feature of the ME-EMSC model, and not merely a consequence of the corresponding peak position in the reference spectrum.

3.5 | Sensitivity toward initialization parameters

As the optimal ranges for the parameters depend on the data set, the following section includes remarks on how the algorithm should be initiated, as well as an assessment on the sensitivity toward the ranges set for the parameters. Prior to correcting a large data set, one should start with a smaller selection of spectra to adjust the parameter settings. In the MATLAB code, this is done in the mode called "PreRun." Correction of the whole data set is done in mode "Correction." Finding the optimal parameters will generally be a process of trial and failure. Descriptions on how the optimal parameter settings are obtained will be related to a concrete example, namely, the correction of the lung cancer cell data set [10].

It is important to note that adjustment of the parameter ranges should be performed without weighting the reference spectrum, or setting negative parts of the reference spectrum to zero. Otherwise, the effects of changing the parameters A, h and the ranges for a and n_0 will become less visible. When the optimal parameter settings are found, the negative parts of the reference spectrum should be set to zero. In the published code, this is automatically handled by selecting the mode ("PreRun" or "Correction"). Weighting is optional after the optimal parameter settings are found, and is turned on by default.

3.5.1 | Setting a, n_0 and h

As a default, the program uses the following physical parameters:

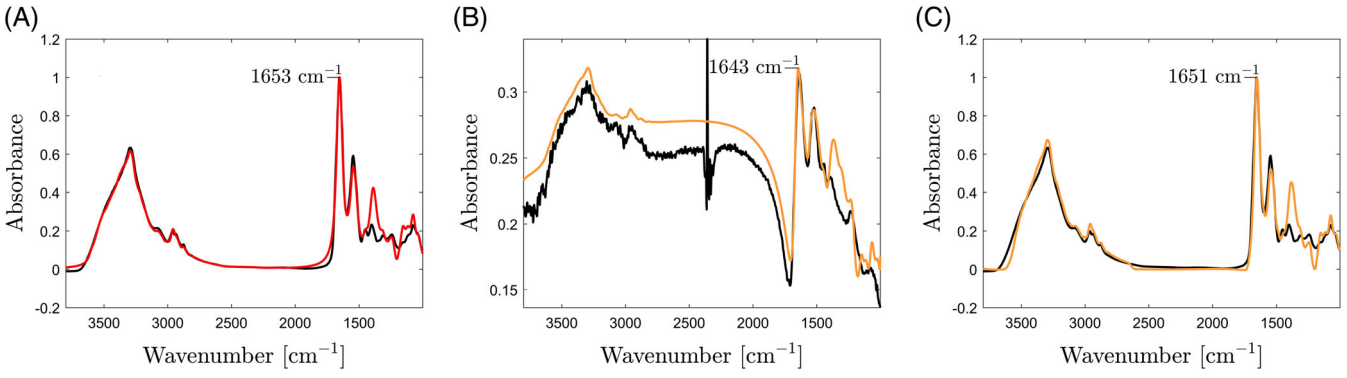


FIGURE 8 A, The simulated pure absorbance spectrum in red, and the Matrigel spectrum in black. B, The amide I peak position is shifted in the simulated apparent absorbance spectra (orange) relative to the corresponding peak position in the underlying pure absorbance spectrum. The measured spectrum used as a template for the scatter contribution is shown in black. This spectrum is obtained from breast cancer cells. C, The corrected simulated apparent absorbance spectrum is shown in orange, and the Matrigel spectrum in black. A more reliable amide I peak position is retrieved with the correction

$a \in [2\ \mu m, 7.1\ \mu m]$.
$n_0 \in [1.1, 1.4]$.
$h = 0.25$.

For the data sets available in this study, the above given parameter ranges result in a stable correction. It is observed that when choosing a parameter range that does not fit to a given data set, the model is not capable of modeling Mie scattering in the measured spectra, and the correction fails. This introduces artifacts in the corrected spectra, where some of the corrected spectra do not resemble absorbance spectra at all. Therefore, it is fairly easy to identify an unsuccessful correction. From the data sets that have been investigated, it is observed that the correction would rather fail than retrieving the wrong chemical information. Artifacts that may occur are usually located in the chemically inactive regions, as well as in the O—H stretching region. The final RMSE of spectra, for which the correction is not successful, is in general significantly higher than for successful corrections. However, it is important to note that the level of acceptable RMSE dependents on the data set, and a visual inspection is recommended until an appropriate parameter range is found. In order to allow adjusting the parameter range for a given data set, the code published together with this paper can be first run in the mode named "PreRun." This mode allows adjusting the parameter ranges for a smaller subset of spectra in the data set. As mentioned, this mode pacifies the down-weighting of the chemically inactive regions in order to allow detecting unsuccessful baseline corrections and visual artifacts that typically are observed in these regions. In the supporting material, examples are given on the result of corrections where the parameter ranges were set outside of the stable regions. This is shown for both experimentally obtained absorbance spectra, and simulated apparent absorbance spectra.

3.5.2 | Setting the number of loadings A

As described in Section 3.1, the number of loadings A is set based on a desired level of explained variance in the Mie extinction curves. A user also has the option to overwrite this criterion by setting A directly. As a default, we propose to use a level of 99.96% explained variance. For the proposed settings for a, n_0 and h given in the previous section, and with the Matrigel spectrum as reference spectrum, this limit results in $A = 7$ components. Figure 9A) shows a correction of the set of lung cancer cells, where the default parameters for a, n_0, h and explained variance are used. It is evident that the Mie oscillations are not represented precisely and the correction is not optimal. Therefore, we need to increase the explained variance in this case. By setting the limit of the explained variance to 99.99%, we see that the Mie oscillations are modeled more precisely (see Figure 9B). The optimal number of components is therefore in this case $A = 9$. Note that if weighting is used while estimating the optimal explained variance, the effect of a too low number of loadings A is less visible. This is shown in

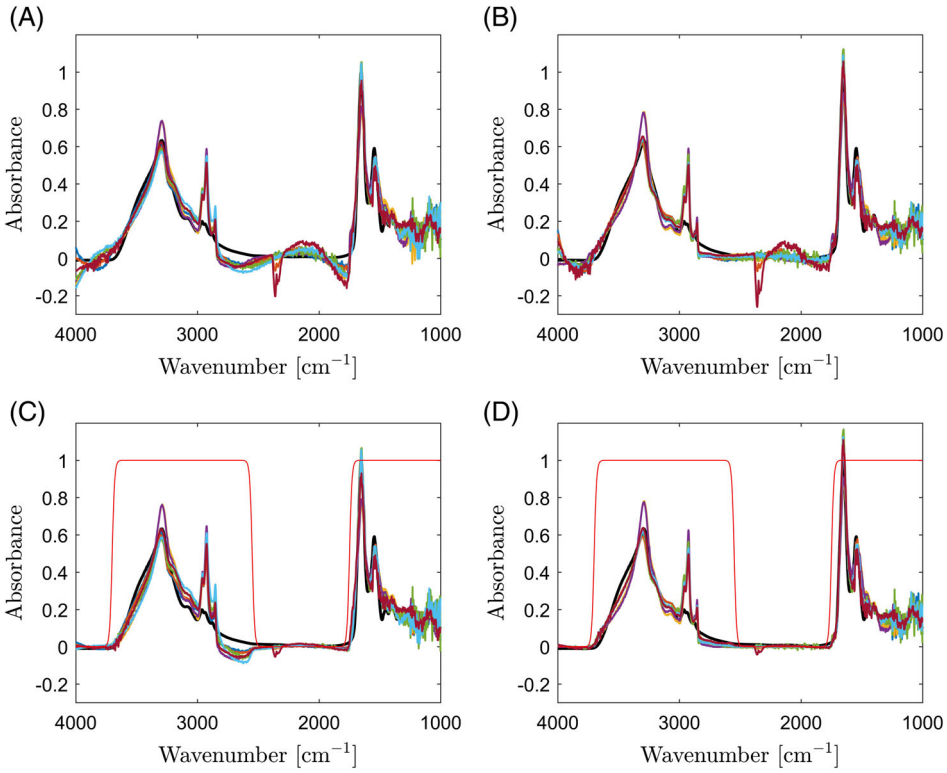


FIGURE 9 Correcting a selection of the lung cancer cell spectra by (A) setting the explained variance in the PCA on the scattering curves to 99.96% and (C) by applying the weight function. B, By setting the explained variance to 99.99%, the baseline correction is better, and D, the weight function can be applied without masking oscillations in the baseline

Figure 9C, where the applied weight function is shown in red.

3.5.3 | Setting the weight function

When applying weighting of the reference spectrum, the user specifies the extension of the chemically active regions, which will correspond to the points of inflection of the hyperbolic tangent function. The following parameters are set as default for the weight function.

Inflection points: 3700, 2550 and 1900 cm^{-1}. Corresponding slope, κ: 1 for all inflection points.

Figure 9D) shows the effect of weighting the reference spectrum with the default weights. By applying the default parameters for the weight function, a smoother baseline correction is achieved for the lung cancer cell data set. In order to adjust the weight function to a data set, the weight function parameters can be adjusted. After the optimal weight function is found, the negative parts of the reference spectrum can be set to zero. In the published MATLAB code, this is done by changing to mode "Correction."

Note that when weighting is used, it requires that the reference spectrum is baseline corrected. If the reference spectrum has a negative baseline, the weighting would result in deformation of the spectrum, which further introduces artifacts in the corrected spectra.

3.5.4 | Correcting the whole data set

Correction can now be performed on the whole data set, with the optimal initialization parameters. This is done in the MATLAB code by changing the mode to "Correction." Failure of the correction for individual spectra can be detected by considering the final RMSE value. The program offers a quality test based on a maximum limit for the RMSE. As default, this limit is set to infinity. As this limit will be dependent on the data set, it should be determined by the user after visual inspection. In the published code, this visual inspection is implemented.

3.5.5 | Speed improvements and run-time analysis

The ME-EMSC algorithm is implemented in MATALB with general considerations to speed optimization. Konevskikh et al. [19] proposed two major improvements for decreasing the computational time, that is, using the Hilbert transform in place of the Fourier transform, and reducing the parameter space of the meta-model. Compared to the two latest versions of the RMieS-EMSC code published by Bassan et al. [9] (version 3 and 5), the new code performs significantly better, with a substantial reduction in computational time, as seen in Figure 10. The ME-EMSC code decreases the runtime with 94% with respect to the RMieS-EMSC code, when correcting spectra of 1428 wavenumbers for 15 iterations per spectrum. The runtime analysis is performed in MATLAB R2018a, on a Lenovo Thinkstation p920 with 20 Intel Xenon cores (2.20 GHz, 126 GB RAM).

An important difference between the RMieS-EMSC algorithm and the ME-EMSC algorithm is the formula used for calculating the Mie scattering. While the RMieS-EMSC algorithm is built on the approximation formula for Mie scattering only, the ME-EMSC also considers the sample absorption, and thereby employs the full Mie extinction approximation formula. As scattering and absorption are mutually dependent on each other, absorption should be taken into account when estimating the scattering from a sample. While using a computationally more expensive approach, we were still able to reduce the computational time significantly compared to the RMieS-EMSC algorithm.

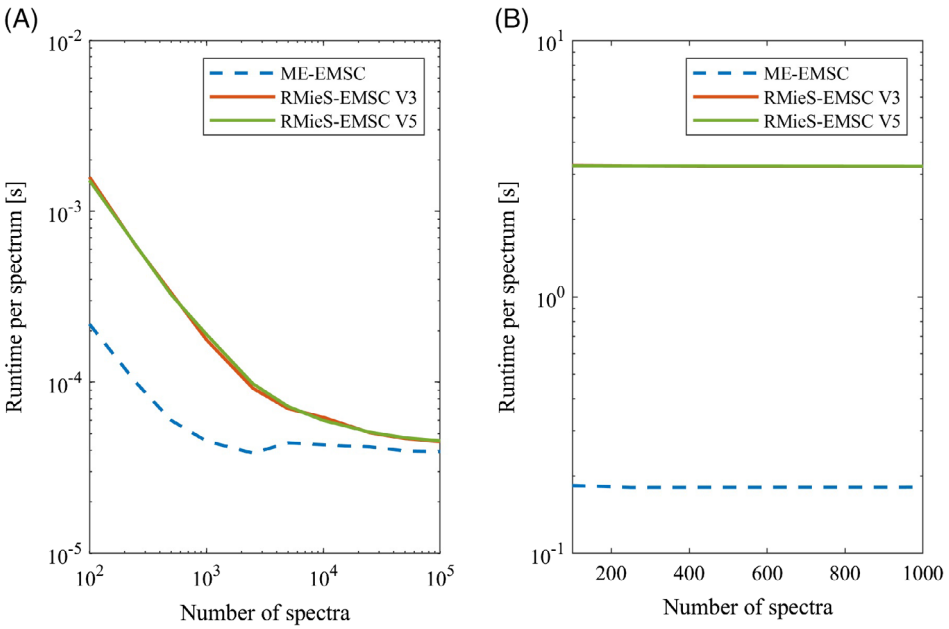


FIGURE 10 Runtime analysis of the improved ME-EMSC algorithm (in dashed blue) and the RMieS-EMSC algorithm of Bassan et al. [8] (version 3 in red and version 5 in green). The runtime is shown per spectrum for (A) 1 iteration per spectrum, and (B) 15 iterations per spectrum

4 | CONCLUSION

In this paper, we presented an optimized code for the correction of scatter-distorted infrared spectra of cells and tissues and published the source code in MATLAB. The presented code is an improvement of earlier published algorithms. We were able to prove that the optimized code overcomes several crucial shortcomings of earlier versions of the code. A major criticism of previous versions of the algorithm was that the corrected spectra acquired features of the reference spectrum that were used to initialize the algorithm. In the paper at hand, we show by using a simulated data set that the latest algorithm allows restoring the true underlying pure absorbance spectrum. Reference spectra with different characteristic features were considered such as a reference spectrum with a shifted peak position a reference spectrum with characteristic features in the O—H stretching region. We could demonstrate that the outcome of the correction is not influenced by the choice of the reference spectrum. In line with this paper, the source code in MATLAB was made accessible at the GitLab repository. In order to provide a stable source code, several stability improvements were done. A weighting of the chemically inactive regions in the reference spectrum for the EMSC parameter estimation and the requirement of non-negativity of the reference spectrum in the iterative algorithm achieved a stable parameter estimation and stable corrections. Improvements of the stop criterion and stable estimates of the number of components make the code user friendly and easy to use on new data sets.

The applicability of the code on simulated spectra and different measured data of cells and tissues proved that the code is universally applicable. Currently, we are studying the application of the Mie correction code on silica beads embedded in a resin matrix with satisfactory results. These results will be published elsewhere. Further work on speed improvements are in progress and updates of the code will be published in the GitLab repository.

ACKNOWLEDGMENTS

The authors would like to thank Tatiana Konevskikh for useful discussions and for providing the fast iterative Mie scatter correction algorithm [19]. We also thank Francisco Peñada, Nick Stone and Ganesh Sockalingum for providing experimental data.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR BIOGRAPHIES

Please see Supporting Information online.

ORCID

Johanne H. Solheim https://orcid.org/0000-0001-6726-1692

REFERENCES

[1] T. Wrobel, R. Bhargava, *Anal. Chem.* **2018**, *90*, 1444.
[2] M. Hermes, R. Brandstrup Morrish, L. Huot, L. Meng, S. Junaid, J. Tomko, G. R. Lloyd, W. T. Masselink, P. Tidemand-Lichtenberg, C. Pedersen, F. Palombo, N. Stone, *J. Opt.* **2018**, *20*, 023002.
[3] R. Bhargava, *Appl. Spectrosc.* **2012**, *66*, 1091.
[4] M. Diem, M. Miljković, B. Bird, T. Chernenko, J. Schubert, E. Marcsisin, A. Mazur, E. Kingston, E. Zuser, K. Papamarkakis, N. Laver, *J. Spectro.* **2012**, *27*, 463.
[5] F. Großerueschkamp, T. Bracht, H. C. Diehl, C. Kuepper, M. Ahrens, A. Kallenbach-Thieltges, A. Mosig, M. Eisenacher, K. Marcus, T. Behrens, T. Brüning, D. Theegarten, B. Sitek and K. Gerwert, *Sci. Rep.* **2017**, *7*, 44829.
[6] C. Kuepper, F. Großerueschkamp, A. Kallenbach-Thieltges, A. Mosig, A. Tannapfel, K. Gerwert, *Faraday Discuss.* **2016**, *187*, 105.
[7] A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel, K. Gerwert, *J. Biophotonics* **2013**, *6*, 88.
[8] P. Bassan, H. J. Byrne, F. Bonnier, J. Lee, P. Dumas, P. Gardner, *Analyst* **2009**, *134*, 1586.
[9] P. Bassan, A. Kohler, H. Martens, J. Lee, H. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, P. Gardner, *Analyst* **2010**, *135*, 268.
[10] A. Kohler, J. Sulé-Suso, G. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. van Pittius, G. Parkes, M. Høy, *Appl. Spectrosc.* **2008**, *62*, 259.
[11] P. Bassan, A. Kohler, H. Martens, J. Lee, E. Jackson, N. Lockyer, P. Dumas, M. Brown, N. Clarke, P. Gardner, *J. Biophotonics* **2010**, *3*, 609.
[12] P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke, P. Gardner, *Analyst* **2012**, *137*, 1370.
[13] T. Van Dijk, D. Mayerich, P. Scott Carney, R. Bhargava, *Appl. Spectrosc.* **2013**, *67*, 546.
[14] R. Lukacs, R. Blümel, B. Zimmermann, M. Bagcioglu, A. Kohler, *Analyst* **2015**, *140*, 3273.
[15] B. Bird, M. Miljković, M. Diem, *J. Biophotonics* **2010**, *3*, 597.
[16] M. Miljković, B. Bird, M. Diem, *Analyst* **2012**, *137*, 3954.
[17] S. B. Banadkoki, F. T. Azar, F. H. Shirazi, *J. Med. Biol. Eng.* **2018**, *1*. https://doi.org/10.1007/s40846-018-0423-9
[18] H. C. van de Hulst, *Light Scattering by Small Particles*, Dover, New York **1981**.
[19] T. Konevskikh, R. Lukacs, R. Blumel, A. Ponossov, A. Kohler, *Faraday Discuss.* **2016**, *187*, 235.
[20] T. Konevskikh, R. Lukacs, A. Kohler, *J. Biophotonics* **2017**, *11*, e201600307.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Solheim JH, Gunko E, Petersen D, Großerüschkamp F, Gerwert K, Kohler A. An open-source code for Mie extinction extended multiplicative signal correction for infrared microscopy spectra of cells and tissues. *J. Biophotonics*. 2019;12:e201800415. https://doi.org/10.1002/jbio.201800415