FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Chemometrics and Statistical Analysis in Raman Spectroscopy-based Biological Investigations

Kumulative Dissertation

zur Erlangung des akademischen Grades Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Chemisch-Geowissenchaftlichen Fakultät
der Friedrich-Schiller-Universität Jena von

M. Eng. Shuxia Guo

geboren am 28.10.1989 in Shanxi, China

1. **Gutachter:** Prof. Dr. Jürgen Popp

   Institut für Physikalische Chemie

   Friedrich-Schiller Universität Jena

2. **Gutachter:** Prof. Dr. Christoph Steinbeck

   Institut für Anorganische und Analytische Chemie

   Friedrich-Schiller Universität Jena

3. **Gutachter:** Prof. Dr. Axel Mosig

   AG Bioinformatik

   Ruhr-Universität Bochum

Tag der öffentlichen Verteidigung:

14 November, 2018

Look up at the stars and not down at your feet.

Try to make sense of what you see,
and wonder about what makes the universe exist.

Be curious.

- Stephen Hawking (1942 - 2018)

Contents

List of Figures **iii**

List of Abbreviations **vi**

1 Introduction **1**

1.1 Raman Spectroscopy 2

1.2 Raman Spectroscopy-based Biological Applications 3

1.3 Chemometrics in Raman Spectroscopy-based Biological Applications 4

2 State of the Art **7**

2.1 Raman Spectral Pre-processing 7

2.1.1 Baseline correction 7

2.1.2 Spectrometer calibration 10

2.2 Statistical Modeling 10

2.3 Sampling in Chemometrics 12

2.4 Prediction of New Data 13

3 Selected Work and Results **15**

3.1 Baseline Correction 17

3.1.1 Automatic optimization of mathematical baseline correction 17

3.1.2 Shifted-excitation Raman difference spectroscopy (SERDS) 21

3.2 Statistical Modeling 25

3.2.1 Model construction and validation 25

3.2.2 Common mistakes in cross-validation 27

3.3 Model Transfer 30

3.3.1 Model transfer approaches 30

3.3.1.1 Data-based model transfer 31

3.3.1.2 Model-based model transfer 31

3.3.2 Validation of model transfer approaches 33

4 Additional Work **35**

5 Summary **37**

6 Zusammenfassung **41**

Bibliography **47**

7 Publications **61**

7.1 Optimization of Raman-Spectrum Baseline Correction in Biological Application (A1) 61

7.2 Spectral Reconstruction for Shifted-Excitation Raman Difference Spectroscopy (SERDS) (A2) 75

7.3 Common Mistakes in Cross-Validating Classification Models (A3) 93

7.4 Towards an Improvement of Model Transferability for Raman Spectroscopy in Biological Applications (A4) 107

7.5 Model Transfer for Raman Spectroscopy based Bacterial Classification (A5) . 121

7.6 EMSC based Model Transfer for Raman Spectroscopy in Biological Applications (A6) 139

Peer-Reviewed Publications **163**

Conferences **165**

Workshops **167**

Acknowledgement **169**

Erklärungen **171**

List of Figures

1.1 An example illustrating the capability of chemometrics to extract subtle spectral differences . . . 5

2.1 Workflow of chemometrics in Raman spectroscopy-based biological applications 8
2.2 Contributions in a measured Raman spectrum . . . 9

3.1 Baseline correction optimized by the marker R^{12} and the Euclidean distance between true and baseline-corrected Raman spectra . . . 18
3.2 Results of R^{12} and classification from the grid search for a three-group task . 19
3.3 Estimated baselines corresponding to the minimal R^{12} and the best classification 20
3.4 Graphic illustration of NNLS-based SERDS spectral reconstruction . . . 22
3.5 Real-world SERDS datasets used for verifying the spectral reconstructionon methods . . . 22
3.6 Reconstuction results of NNLS-based method on the three real-world SERDS datasets . . . 23
3.7 Reconstruction results of antiD-based method on the three real-world SERDS datasets . . . 23
3.8 Reconstruction results of FT-based method on the three real-world SERDS datasets . . . 23
3.9 Framework of statistical modeling using two-layer cross-validation for model optimization and evaluation . . . 26
3.10 Results of testing accuracy from the external cross-validation for the two binary classification tasks . . . 27
3.11 Results of the testing accuracies and the validation accuracies in different cases of internal cross-validation . . . 28
3.12 Prediction of the secondary dataset in different cases of model transfer . . . 34
3.13 Comparison between secondary prediction and primary prediction using different model transfer methods . . . 34

List of Abbreviations

antiD	anti-derivative
nLV	number of latent variables
nPC	number of principal components
ALS	asymmetric least squares
CV	cross-validation
ED	Euclidean distance
EMSC	extended multiplicative signal correction
FWHM	full-width-at-half-maximum
FT	Fourier transform
LDA	linear discriminant analysis
Modpoly	modified polynomial
MS	score movement of principal component analysis
NNLS	non-negative least squares
PCA	principal component analysis
PLS	partial least squares
SA	spectral augmentation
SERDS	shifted-excitation Raman difference spectroscopy
SNIP	sensitive nonlinear iterative peak clipping
SNR	signal-to-noise ratio
SVM	support vector machine
TR	Tikhonov regularization

1. Introduction

Over the last two decades, the health care systems have encountered new challenges and experienced severe pressure, mainly because of three factors. As the first factor, the spread of infectious diseases has been dramatically sped up by the expanding global transport network [1]. Additionally, chronic and lifestyle-related diseases appear to have increased in prevalence, particularly due to the diet and lifestyle changes in industrial societies [2,3]. More importantly, age-related diseases have become a major problem in developed countries as a result of the rapid demographic change towards an aging society [4]. All three of these factors make it urgently necessary to improve both disease care and individual health maintenance. This improvement requires deeper knowledge about the origin and progression of diseases, especially the understanding of chemical and biochemical processes on a molecular level [5]. The starting point is to capture signatures of chemical and biochemical components from biological samples. One of the approaches to doing so is molecular imaging.

Molecular imaging has enabled the visualization of biological processes in living organisms on a cellular and molecular level [5]. Existing molecular imaging techniques include, but are not limited to, positron emission tomography (PET), single photon-emission computed tomography (SPECT), magnetic resonance imaging (MRI), and optical imaging [6]. PET and SPECT can provide accurate quantitative information on molecules with high sensitivity. However, their potential is restricted by the relatively low spatial resolution, high cost, and risk of radiation exposure. MRI has gained attention in clinical use due to its excellent spatial resolution, but it is limited by low sensitivity [7, 8]. Optical imaging is emerging as a more important molecular imaging technique with great potential in real-time measurement, quantitative analysis, multiplexing, and endoscopy [9]. In this technique, molecular signatures are detected based on the optical properties of samples, such as absorption, scattering, polarization, spectral characteristics, and fluorescence [10]. Typical examples of optical imaging include fluorescence imaging, infrared (IR) imaging, and Raman imaging [5, 6, 11].

A large variety of biomolecules can be detected by fluorescence imaging, thanks to the continuous discovery of fluorescent probes [12]. With elaborate techniques, fluorescence imaging is able to provide sub-diffraction spatial resolution [10, 13–15], to trace dynamic processes within cells [16–19], and to be applied *in vivo* [20, 21]. Like all other labeling imaging techniques, however, fluorescence imaging is hindered by the fluorescence probes required. Information that can be accessed by fluorescence imaging is limited by the access of probes, and it can be a very extensive process to develop novel probes. Even worse, fluorescence probes may disturb the properties of biological molecules or be toxic to human subjects. Additionally, the reproducibility of probes and labeling procedures largely influences the quality of measured data [5, 22, 23].

Therefore, label-free optical techniques, such as infrared (IR) absorption and Raman spectroscopy, are more attractive. Both techniques provide highly-specific molecular fingerprint information and characterize the molecular environment in biological samples. However, IR spectroscopy is often not applicable in biological applications due to strong water absorption [22]. In contrast, Raman spectroscopy is insensitive to water and ideally suitable for biological investigations [24]. To understand the molecular fingerprints of biological samples represented by Raman spectroscopy, it is necessary to briefly refer to the theory of Raman spectroscopy.

1.1 Raman Spectroscopy

Raman spectroscopy is a spectroscopic technique based on Raman scattering, a phenomenon discovered by C.V. Raman [25, 26] and G. Landsberg [27] in 1928. The theory of Raman scattering can be explained by investigating the interaction between light and molecules [5, 28]. The oscillating electric field $E = E_0 cos(\omega_0 t)$ of incoming light can displace the electron cloud of a molecule against the atomic nuclei and induce a dipole moment within the molecule. The strength of this induced dipole moment is shown by:

$$\mu = \alpha \cdot E = \alpha \cdot E_0 cos(\omega_0 t), \tag{1.1}$$

where the term α denotes the polarizability measuring how easily the electrons can be distorted within a molecule. The polarizability is not constant but can change because of molecular vibration. The dependence of the polarizability on the nuclear coordinate q can be expressed by a Taylor series expanded around the equilibrium nuclear geometry $q = 0$:

$$\alpha = \alpha(q) = \alpha(0) + (\frac{\partial \alpha}{\partial q})\Big|_{q=0} \cdot q + \cdots . \tag{1.2}$$

A molecular vibration with the characteristic frequency ω_R can be approximated as a harmonic oscillation around the equilibrium nuclear geometry:

$$q = q_0 \cdot cos(\omega_R t). \tag{1.3}$$

Substituting Eq. (1.2-1.3) into Eq. (1.1) and omitting the nonlinear terms in Eq. (1.2), the induced dipole moment is formulated by:

$$\mu = \left[\alpha(0) + (\frac{\partial \alpha}{\partial q})\Big|_{q=0} \cdot q_0 \cdot cos(\omega_R t) \right] \cdot E_0 cos(\omega_0 t). \tag{1.4}$$

Eq. (1.4) can be rewritten as:

$$\begin{aligned}\mu = \alpha(0) \cdot E_0 cos(\omega_0 t) \quad &+ \quad \frac{1}{2}\,(\frac{\partial\alpha}{\partial q})\bigg|_{q=0} \cdot q_0 \cdot cos((\omega_0 - \omega_R)t) \\ &+ \quad \frac{1}{2}\,(\frac{\partial\alpha}{\partial q})\bigg|_{q=0} \cdot q_0 \cdot cos((\omega_0 + \omega_R)t). \end{aligned} \tag{1.5}$$

Eq. (1.5) shows that the induced dipole oscillates at three frequencies. The first frequency is equal to the excitation frequency ω_0. The other two frequencies, $\omega_0 - \omega_R$ and $\omega_0 + \omega_R$, are the difference or the sum of the excitation frequency ω_0 and the characteristic frequency ω_R of the molecular vibration. The induced dipole, in turn, radiates scattered light at these three frequencies. The radiation with frequency ω_0 is called Rayleigh scattering. The radiation with frequencies $\omega_0 - \omega_R$ and $\omega_0 + \omega_R$ is referred to as Stokes-Raman scattering and anti-Stokes Raman scattering, respectively. It can also be derived from Eq. (1.5) that Raman scattering only occurs if the polarizability α is changed by the molecular vibration (i.e., $\frac{\partial\alpha}{\partial q} \neq 0$). Molecules showing Raman response are termed Raman-active molecules. The Raman scattering of a Raman-active molecule represents the vibrational fingerprints of this molecule because the frequency shift of the Raman scattering relative to the excitation frequency is determined by the characteristic frequency ω_R of the molecular vibration [29,30]. It is thus possible to identify all Raman-active molecules contained in a sample by analyzing Raman spectra of this sample.

1.2 Raman Spectroscopy-based Biological Applications

The previous section shows in theory how Raman spectra contain vibrational fingerprints of Raman-active molecules. Since most biomolecules are Raman-active, rich molecular fingerprints of biological samples can be delivered via Raman spectra. Moreover, Raman spectroscopic measurements of biological samples are not disturbed by water considering the low Raman response of hydroxyl groups [24]. Hence, Raman spectroscopy is highly suitable for measurements of biological samples.

The early biological application of Raman spectroscopy was reported in 1936 [31]. However, it took almost 40 years before Raman spectroscopy became popular in biological research [30, 32]. This slow development was the result of the intrinsically small cross-section of Raman scattering and the low concentration of biomolecules [5]. It required long integration times to obtain a utilizable signal. This situation changed with the invention of a laser that has higher excitation intensities [33]. Further, it became possible to acquire large datasets with a wide wavenumber range thanks to the technical advances of effective Rayleigh filters [34, 35], computers, and low-noise detectors such as a charge-coupled device (CCD) [36, 37]. As a result, Raman spectroscopy could be successfully employed to study biomolecules such as proteins, nucleic acids, and lipids [38, 39] in the 1970s. Later, in 1976, Dick Lord reported the first-version strategy and tactics used to apply Raman spectroscopy for biomolecules at the FACSS meeting in Philadelphia [40]. Recently, Butler and coauthors published a standard protocol for employing Raman spectroscopy to study biological materials from the perspectives of instrumentation, spectral acquisition, sample preparation, and data analysis [37].

Nowadays, Raman spectroscopy is widely applied in modern biological applications [11, 41–43]. There are numerous studies available covering various fields, such as toxicology [44, 45], microbiology [46–48], drug discovery [49–51], metabolic investigations [52–55], and forensic analysis [56, 57]. Raman spectroscopy can also be applied *in vivo* and thus benefits medical diagnostics and therapeutic interventions [58–63]. The most illustrative example is the detection of early-state cancers, such as cancers of the skin [64, 65], mouth [66, 67], brain [68–70], larynx [71, 72], breast [73, 74], lungs [75, 76], lymph nodes [77, 78], bladder [79, 80], colon [81, 82], prostate [83], uterus [84], and so on. Raman spectroscopic applications for the detection of other diseases have been reported as well. Examples of such applications include Raman spectroscopy-based studies of inflammatory bowel diseases [85] and Alzheimer's disease [86].

Nonetheless, the above-mentioned applications are only possible if the spectral signal is effectively translated into high-level information such as disease levels [42]. The translation is far from straightforward because the measured Raman spectra are often contaminated by corrupting effects such as fluorescence emission. Another significant difficulty in the translation is that the spectral variations caused by biological changes are very subtle and cannot be detected by the naked eye or a simple database search. It is required to incorporate more advanced analysis approaches into Raman spectroscopy-based techniques in order to remove the corrupting effects and detect the subtle spectral variations of interest [5, 43]. This leads to the topic of chemometrics [87]. The general idea of chemometrics and its close connection with Raman spectroscopy is described in the following section.

1.3 Chemometrics in Raman Spectroscopy-based Biological Applications

The term 'chemometrics' was first used in 1972 [88], ten years after the first application of multivariate methods in chemistry to determine the number of components for the fluorescence spectra of mixtures [89]. It took another decade before chemometrics became widely used, thanks to the NATO-sponsored meeting held in Cosenza, Italy [90]. Since then, chemometrics has been a cornerstone of analytical chemistry. The basic idea of chemometrics is simple. A statistical model is built on a certain number of known samples, namely: the training data. This trained model is then validated according to its prediction on a dataset that is independent on the training data. The independent dataset is termed testing data.

Finally, the model is saved and used for predicting unknown samples in the future.

Chemometrics plays an essential role in Raman spectroscopy-based biological investigations, which has been revealed from multiple aspects [5, 11, 42]. Foremost, a manual spectral inspection requires an expert's interaction and is subjective. The same expert can make controversial conclusions on the same sample at different times. On the contrary, chemometric approaches do not require human interaction and can provide consistent and objective output. Further, corrupting signals contained in Raman spectra, such as fluorescence emission, can be removed by chemometric methods, leading to an improved data quality and better prediction performance. Additionally, the datasets in biological applications are extensive and impossible to be handled manually. Chemometric approaches make it possible to handle a massive amount of data very fast, thanks to the powerful computation capabilities of modern computers.

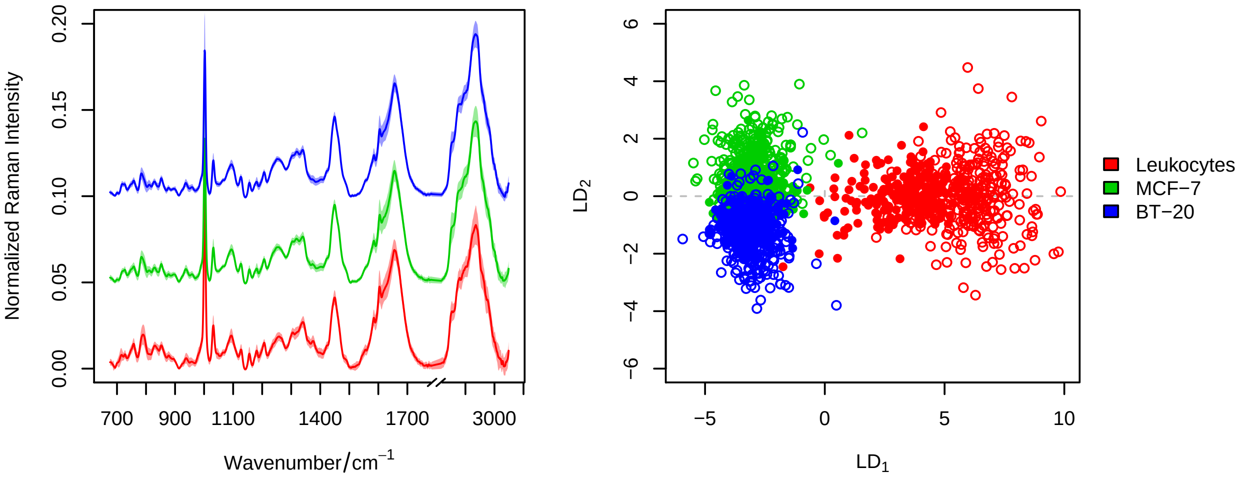


Figure 1.1: An example illustrating the capability of chemometrics to extract subtle spectral differences. The three cell types (leukocytes, MCF-7, and BT-20) are difficult to distinguish by the naked eye according to the mean Raman spectra shown in the left panel. From the LD scores calculated from the chemometric model (PCA-LDA), as are shown in the right panel, the three cell types become clearly differentiable. The open and filled circles in the right panel represent the scores of training (six out of nine replicates) and testing (the rest three replicates) data, respectively.

More importantly, chemometric approaches can extract the subtle spectral variations caused by biological changes of interest, which enhances the sensitivity of Raman spectroscopy-based biological detection. This enhancement is proven by numerous studies [46, 62, 70, 78, 91–95]. An illustrative example is shown in Figure 1.1. In this example, the Raman spectra of leukocytes and breast carcinoma-derived tumor cells (MCF-7 and BT-20) were measured. Each cell type consists of nine replicates [92]. The three cell types are difficult to distinguish by the naked eye according to the mean spectra shown in the left panel of Figure 1.1. By using a statistical model, which is composed of principal component analysis (PCA) and linear discriminant analysis (LDA), the three cell types can be easily distinguished, as shown in the right panel of Figure 1.1. The clear separation of the three cell types shows that the chemometric model can successfully extract the subtle spectral differences of the three cell types.

Nonetheless, there are still unresolved issues in the application of chemometrics in Raman spectroscopy-based biological investigations. Chemometric techniques are required to automatically optimize the baseline correction [96,97], to effectively remove the extremely intense fluorescence background [98], to reliably optimize and evaluate a statistical model [99, 100], and to adequately predict data measured from a new replicate with the trained model [101]. These open issues in chemometrics motivated the investigations in this thesis. Detailed explanations of each topic, as well as the proposed approaches, will be provided in chapter 3, following an overviewed state of the art of chemometric techniques in chapter 2.

2. State of the Art

Before presenting the work related to the aforementioned open issues, it is necessary to overview the state of the art of chemometric techniques. A typical workflow of chemometrics in Raman spectroscopy-based biological applications is illustrated in Figure 2.1. According to this workflow, the overview is divided into four aspects: Raman spectral pre-processing, statistical modeling, sampling, and prediction of new data.

2.1 Raman Spectral Pre-processing

Chemometrics in Raman spectroscopy often starts from Raman spectral pre-processing. The reason for this is that measured Raman spectra are often contaminated by corrupting effects such as cosmic spikes, fluorescence, and Gaussian and Poisson noise (Figure 2.2) [96]. They can hamper subtle spectral variations caused by biological changes of interest and lead to decreased performance of the subsequent analysis. The removal of these contaminations is necessary and usually achieved by Raman spectral pre-processing steps including de-spiking, spectrometer calibration, baseline correction, smoothing, outlier detection, and normalization [42,96,102,103], as shown in Figure 2.1. The following text of this subsection will focus on baseline correction and spectrometer calibration. Other steps are beyond the scope of this thesis; details can be found in the references [42,96,104,105].

2.1.1 Baseline correction

Fluorescence emission manifests as a slowly changing baseline profile under Raman bands. It is one of the most influential corrupting effects in measured Raman spectra because it can be several orders of magnitude more intense than Raman scattering. The removal of such fluorescence profile is the central aim of baseline correction. Existing baseline correction methods are based on two mechanisms: mathematical baseline correction [106–109] and experimental baseline correction [110,111]. A brief overview of these two baseline correction categories is provided in the following.

With mathematical baseline correction, the fluorescence baseline is estimated mathematically and subtracted from the acquired Raman spectrum. Methods like polynomial

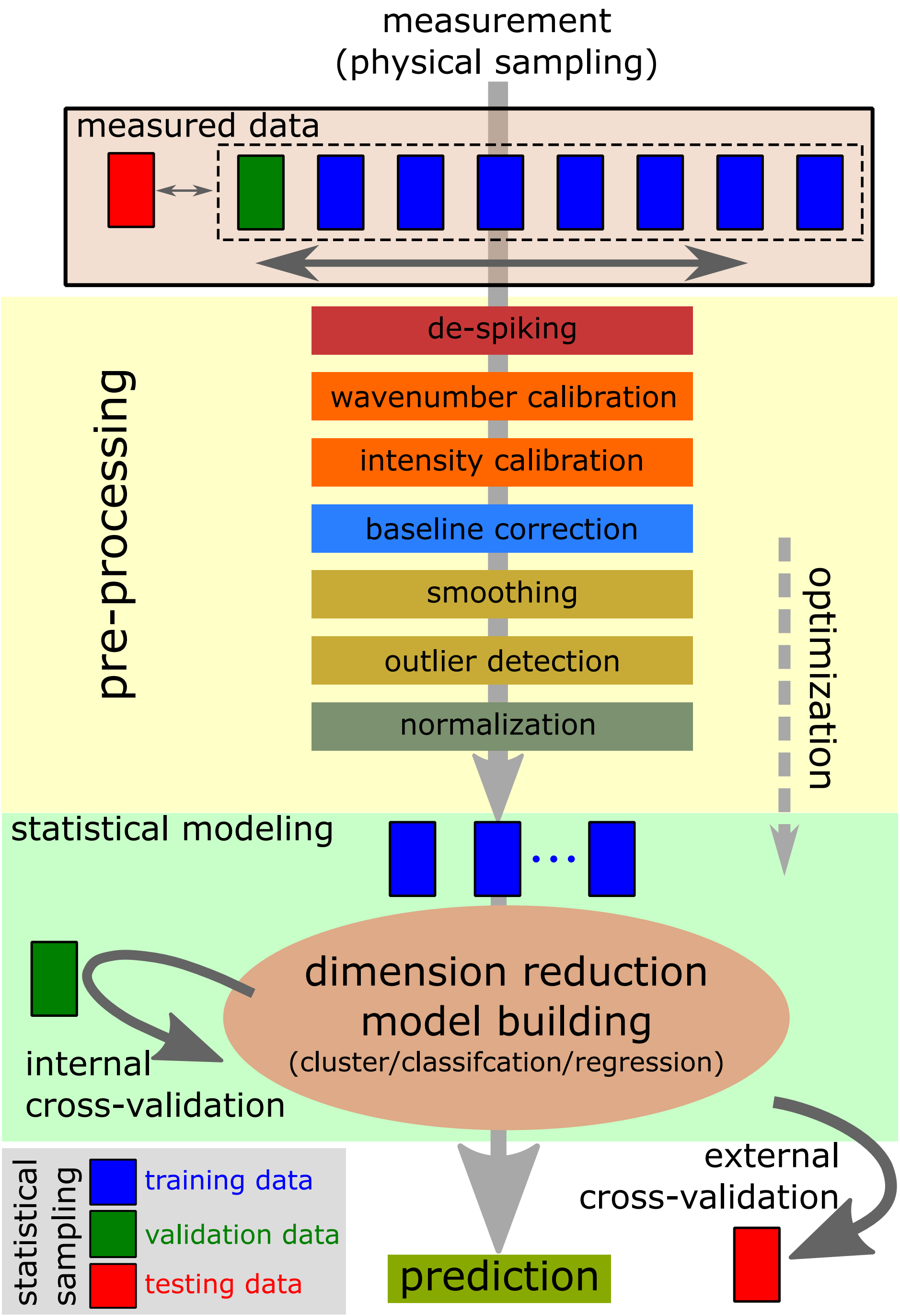


Figure 2.1: Workflow of chemometrics in Raman spectroscopy-based biological applications: The pre-processing removes corrupting effects within a measured Raman spectrum. Statistical modeling translates the spectral signal into high-level information, such as disease levels, which usually consists of dimension reduction and model building. Two-layer cross-validation (CV), composed of internal and external CV, is often necessary in order to achieve both model optimization and evaluation. The model parameters are optimized with internal CV, while external CV is used for model evaluations. Pre-processing parameters can also be optimized by including the step to be optimized inside the internal CV loop.

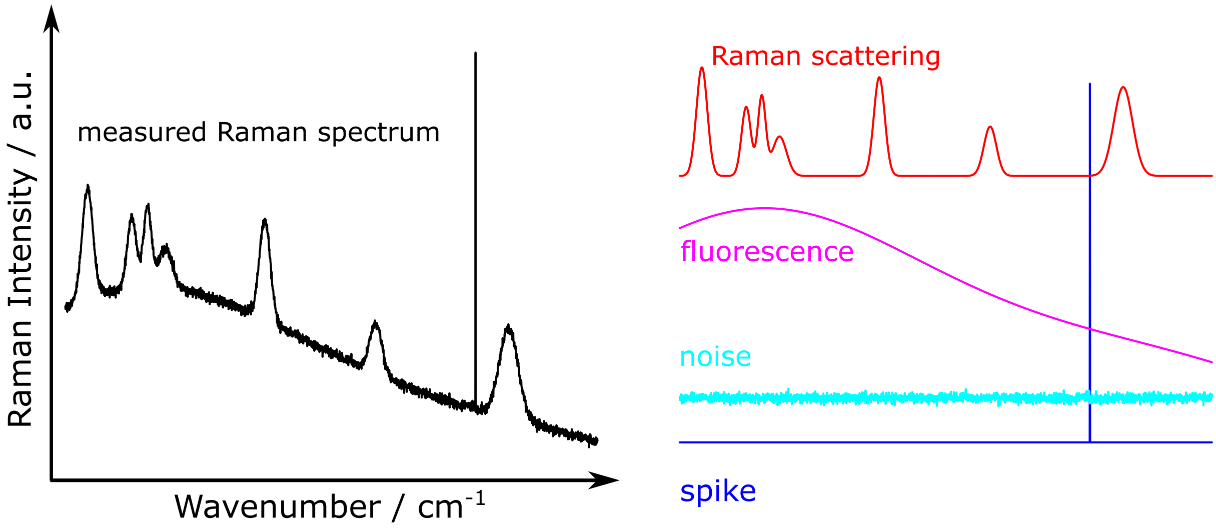


Figure 2.2: Contributions in a measured Raman spectrum: In addition to Raman signals, a measured Raman spectrum is corrupted by cosmic spikes, the fluorescence baseline, and noise. Raman spectral pre-processing is necessary to remove these corrupting components.

fitting [107], asymmetric least squares (ALS) [108], and extended multiplicative signal correction (EMSC) [109] fall into this category. Mathematical baseline correction is most widely applied due to its flexibility and low cost. However, without carefully optimizing the parameters, the baseline correction might lead to a significant loss of useful spectral information and decrease in the performance of subsequent analysis [96,97,112]. A manual parameter adjustment is possible but time-consuming and subjective, especially for biological investigations with a massive amount of data. Automatic optimization has been achieved by model-based methods, in which the baseline correction was optimized by seeking for the best performance of the subsequent model such as a classifier [96,97,112]. The major drawback of these optimization methods is that they require a large amount of training samples to build the subsequent model. The result of such optimization is also dependent on the employed model. A different procedure is necessary to realize a model-independent optimization, which will be addressed in subsection 3.1.1.

With experimental baseline correction, the fluorescence baseline is removed via technical modification of the instruments. The related methods include Raman spectroscopy with near-infrared (NIR) excitation [110], time-resolved Raman spectroscopy [111], and polarization-resolved Raman spectroscopy [113]. All of these methods have certain limitations, and the required instrumental modification can be very expensive [114]. An alternative method is shifted-excitation Raman difference spectroscopy (SERDS), in which two Raman spectra are measured at two slightly different excitation wavelengths to obtain a difference spectrum without fluorescence [98, 115–117]. However, the difference spectrum does not directly show the Raman bands and is difficult to interpret. It is required to reconstruct a fluorescence-free Raman spectrum from the two recorded Raman spectra. The existing methods for such spectral reconstruction are based on numerical peak fitting, anti-derivative, or Fourier transform [98, 118–121], of which all suffer from certain limitations. The numerical peak fitting is inapplicable for severely overlapping Raman bands, as are often observed in biological applications. Spectral resolution is significantly degraded by the anti-derivative-based reconstruction. The Fourier transform-based method is hampered by the high-frequency artifacts due to the frequency leakage. In addition, reconstruction with these methods is largely hindered by intensity variations between the two recorded Raman spectra caused by unavoidable experimental changes. A new spectral reconstruction method is required and proposed in subsection 3.1.2.

2.1.2 Spectrometer calibration

Another significant step of Raman spectral pre-processing is the spectrometer calibration, which is performed to remove the influence of the response of an instrument on a measured Raman spectrum. Because of the influence of the instruments, a measured Raman spectrum can be significantly different from the true values in both wavenumber and intensity axes. These spectral deviations can be removed by a spectrometer calibration: the response function of an instrument is calculated, and observed Raman shifts and intensities are related to their true values according to the calculated response function. The procedure includes wavenumber calibration and intensity calibration [28, 122, 123].

The wavenumber calibration requires a reference Raman spectrum to be measured from a standard material that has well-defined Raman bands [124]. The difference between the measured and theoretical band positions of these known Raman bands is calculated and used to fit a parametric function. This parametric function, which represents the relationship between the observed and the true wavenumber axis, is then employed to correct the wavenumber axis of measured Raman spectra.

The intensity calibration requires the measurement of a standard material with a known emission at different frequencies [28]. The ratio between the measured and theoretical emission of this standard material is calculated, which is the intensity response function of the instrument. The intensities of a measured Raman spectrum are divided by this intensity response function for (linear) intensity calibration.

2.2 Statistical Modeling

After pre-processing, the corrected Raman spectra are used for statistical modeling to translate spectral signals into high-level information like disease levels. Statistical modeling normally starts with a dimension reduction procedure, which helps not only to decrease computational effort but also to improve the generalization performance of a model. Dimension reduction can be achieved in a supervised or unsupervised manner, such as with partial least squares (PLS) and principal component analysis (PCA), respectively. Their difference is whether or not the response variables are used during computation. Dimension reduction methods can also be categorized into feature extraction and feature selection [125]. Feature extraction methods transform the original data space into a new coordinate system with a lower dimension. Typical examples include PCA [87], PLS [126], independent component analysis (ICA) [127], and multivariate curve resolution alternating least squares (MCR-ALS) [128]. In contrast, feature selection approaches pick 'important' variables according to a chosen metric. Methods falling into this category include competitive adaptive reweighted sampling [129], feature selection based on Fisher's discriminant ratio (FDR) [130], Monte Carlo based methods [131, 132], and many others [133, 134]. Over the last two decades, dimension reduction has been facilitated by advanced algorithms (e.g., Isomap [135], locally linear embedding (LLE) [136, 137], auto encoder, and other neural network-based techniques [138, 139]).

After dimension reduction, the lower-dimensional data is fed into a statistical model [42], be it clustering, classification, or regression. Frequently applied algorithms include k-means [140] and hierarchical clustering analysis (HCA) [141] for clustering, linear discriminant analysis (LDA) and random forest (RF) [142, 143] for classification, and principal component regression (PCR) [126] for regression. Moreover, intrinsic regression methods, such as PLS, support vector machine (SVM), and artificial neural network (ANN), are applicable in classification as well by codifying the class information as a dummy response variable (0 or 1) [93, 144, 145]. In particular, the statistical modeling can be conducted in a hierarchical manner for multi-group tasks, as performed in the reference [94]. This makes it possible to incorporate the biological information of samples into statistical modeling.

Statistical modeling has seen new developments from multiple perspectives. First, sparse and fuzzy extensions of existing models are utilized to achieve faster computation, higher stability, and better generalization [146]. Second, fusion techniques, including data fusion (multi-block, multi-group analysis) [147], model fusion (decision trees, classifier ensemble) [148], and decision fusion [149], have found their place to build a more powerful model or to obtain a more stable evaluation of model performance. Third, local modeling has been reported to tackle nonlinear problems more effectively [150]. Further advances are possible with deep learning technologies, which can achieve nonlinear feature extraction and classification [151, 152].

Regardless of the methods applied, a common issue in statistical modeling is over-fitting. It means the model fits training data too perfectly and cannot be generalized to new samples. A common way to avoid over-fitting is to optimize the model by minimizing the prediction error on a dataset different from the training data, namely: validation data. This optimization ensures a trade-off between the training error and the generalization performance of the model [153]. After the optimization, the optimized model also needs to be evaluated to estimate its performance in predicting new data. This is often carried out by using the model to predict a dataset that is independent on the training and validation data. This independent dataset is referred to as testing data. The model optimization and evaluation are extremely important to ensure a high accuracy and robustness of Raman spectroscopy-based biological detection. These two procedures can be conducted under a framework of two-layer cross-validation (CV) (see Figure 2.1) [154], which is composed of internal and external CV. Details of two-layer CV are outlined in subsection 3.2.1.

2.3 Sampling in Chemometrics

Aside from the aforementioned procedures, a hard prerequisite of chemometrics is that the measured samples are good representatives of the population of interest. The related topics fall into the field of design of experiment (DoE), which was first introduced into modern statistical concepts by Fisher [155]. The central aspect of DoE is the theory of sampling (TOS), which has two meanings in chemometrics: physical sampling and statistical sampling [156, 157].

Physical sampling occurs prior to data acquisition. The key is to design a proper protocol so that the samples to be measured represent the population of interest well. Improper physical sampling was reported to increase the error of statistical analysis by 10 $\sim$ 1000 times [156]. One of the most critical parameters for proper physical sampling is the sample size. The related topic, sample size planning (SSP), has been investigated in several studies [158–160]. An extremely significant issue in SSP is that the properties of the samples have to be well considered during the planning, especially in biological investigations. This is because the biological experiment cannot be totally controlled and there are always unknown variations in replicates or individuals. Therefore, the sample size should be counted as the number of replicates/individuals rather than the number of measured spectra [160].

Statistical sampling (or resampling) refers to the procedure of drawing a subset from an available dataset, which is often used in chemometrics for parameter estimation or model validation [161, 162]. The commonly applied resampling methods include Jacknife, hold-out, bootstrapping, and cross-validation. Details of these techniques are provided in references [163, 164]. Similar to the case of SSP, proper resampling has to take the distribution of the population into account. This point can never be over-stated when it comes to model evaluation for Raman spectroscopy-based biological applications, in which new samples to be predicted are often significantly different from the training data due to the inter-replicate variations. By random resampling, the testing data used for model evaluation is not independent on the training data. The difference between such testing and training data does not represent the difference between the training and new data from real-world applications. The model evaluation in this case does not truly reflect the performance of a model in future prediction. Therefore, the resampling has to be conducted at the highest hierarchical level of samples instead of a random manner. If the dataset is composed of a large number of spectra from multiple individuals, the spectra belonging to the same individual should be treated altogether as ‘one’ data package [162, 163, 165, 166]. The influence of improper resampling on model evaluation is unraveled in subsection 3.2.2.

2.4 Prediction of New Data

After a model is constructed and evaluated, it is saved and used to predict new data in order to achieve tasks like disease diagnosis. Despite reliable model optimization and evaluation, the prediction for new data is mostly worse than the prediction for training data. This phenomenon is known as the shrinkage effect of predictors [167], caused by the unavoidable difference between new and training data. The shrinkage effect can be very severe in Raman spectroscopy-based biological investigations, in which difference between new data and training data can be comparable or even larger than subtle spectral variations originating from biological changes of interest. Such undesirable spectral difference can be caused by inter-replicate or inter-individual variations, as well as instrument-related changes. Instrument-related spectral changes can be reduced by the above-mentioned spectrometer calibration. However, the spectrometer calibration is incapable of reducing spectral variations caused by inter-replicate variations. As a result, new data from a replicate/individual different from the training data is impossible to be adequately predicted by the pre-trained model. One solution is to rebuild a specific model for this new data. However, this requires a large number of new training samples, which can be expensive or even impossible to measure.

An alternative technique of handling this failed prediction is model transfer, which aims to enable the trained model to successfully predict the new data [101, 168]. Existing methods for model transfer include procrustes analysis [169], piece-wise direct standardization (PDS) [169], warping [170], global modeling [101], sample-wise spectral multivariate calibration [171], model augmentation [172], and so on. These approaches are mostly applied in near-infrared spectroscopy and regression problems [101, 168]. Model transfer of Raman spectroscopy and classification tasks is scarce [173] and required to be developed. Details of this topic and the proposed approaches are described in section 3.3.

3. Selected Work and Results

This chapter presents the investigations and approaches to filling the gaps of applying chemometrics in Raman spectroscopy-based biological studies, as were described in the last paragraph of chapter 1. The related topics are briefly outlined in the following:

- Fluorescence emission in a measured Raman spectrum can be removed by a mathematical baseline correction method, of which the parameters have to be carefully optimized to preserve the useful spectral signals. In biological applications, such an optimization must be done automatically in order to handle the massive datasets effectively. To do so, a quantitative marker is defined in subsection 3.1.1 as a figure-of-merit of baseline correction. An automatic baseline correction optimization procedure is established based on the defined marker.

- Mathematical baseline correction is inapplicable if the fluorescence is too intense and masks the Raman signal. The shifted-excitation Raman difference spectroscopy (SERDS) can be employed as an alternative option, in which two Raman spectra are measured with slightly different excitation wavelengths. A successful application of SERDS requires to reconstruct the fluorescence-free Raman spectrum from the two measured Raman spectra. Such spectral reconstruction is achieved by non-negative least squares (NNLS)-based method in subsection 3.1.2.

- Model optimization and evaluation are two essential procedures in statistical modeling. The model optimization helps to avoid over-fitting by seeking for a trade-off between the training error and the generalization performance of a model. The model evaluation is used to estimate the performance of an optimized model in predicting new data. Proper model optimization and evaluation are extremely important to ensure a high accuracy and robustness of Raman spectroscopy-based biological diagnostics. Therefore, a guideline is necessary and proposed in section 3.2 showing how to reliably optimize and evaluate a statistical model.

- A pre-trained model can be used to predict unknown (new) data in order to achieve tasks such as disease diagnosis in Raman spectroscopy-based biological applications. The prediction can fail if new data is significantly different from training data due to the inter-replicate spectral variations. One possible solution is to rebuild a specific model for the new data. However, this requires a large number of new training samples, which can be expensive or impossible to measure. An alternative strategy is model transfer, to which the related approaches are developed and described in section 3.3.

The descriptions and discussions in the next sections will be based on the following publications and manuscripts (in the order of their appearance in the text, reprints are provided in chapter 7):

[A1] S. Guo, T. Bocklitz, and J. Popp
Optimization of Raman-spectrum baseline correction in biological application
Analyst, 2016, 141, 2396-2404.

[A2] S. Guo, O. Chernavskaia, J. Popp, and T. Bocklitz
Spectral reconstruction for shifted excitation Raman difference spectroscopy (SERDS)
Talanta, 2018, 186, 372-380.

[A3] S. Guo, T. Bocklitz, U. Neugebauer, and J. Popp
Common mistakes in cross-validating classification models
Analytical Methods, 2017, 9, 4410-4417.

[A4] S. Guo, R. Heinke, S. Stöckel, P. Rösch, T. Bocklitz, and J. Popp
Towards an improvement of model transferability for Raman spectroscopy in biological applications
Vibrational Spectroscopy, 2017, 91, 111-118.

[A5] S. Guo, R. Heinke, S. Stöckel, P. Rösch, J. Popp, and T. Bocklitz
Model transfer for Raman spectroscopy based bacterial classification
Journal of Raman Spectroscopy, 2018, 49, 627-637.

[A6] S. Guo, A. Kohler, B. Zimmermann, R. Heinke, S. Stöckel, P. Rösch, J. Popp, and T. Bocklitz
EMSC based model transfer for Raman spectroscopy in biological applications
Analytical Chemistry, submitted.

3.1 Baseline Correction

As mentioned previously, the fluorescence baseline is one of the most influential corrupting effects in a measured Raman spectrum and has to be removed prior to statistical modeling. Existing approaches to removing this baseline can be categorized as mathematical baseline correction and experimental baseline correction. The issues related to the two mechanisms are investigated in the following subsections.

3.1.1 Automatic optimization of mathematical baseline correction

Mathematical baseline correction is widely applied due to its low cost and high flexibility. However, improper baseline correction can degrade the quality of subsequent analysis. It is important to carefully optimize the methods and the parameters for each specific dataset to ensure a reasonable baseline correction [96,97,112]. Such optimization needs to be automatic so that a massive amount of data in biological investigations can be handled effectively. The approach to doing so is developed and presented in this subsection. The related work was published in [A1].

The proposed optimization method is based on a quantitative marker R^{12} defined as a figure-of-merit of baseline correction. The definition of R^{12} is shown by:

$$\begin{aligned} m_1 &= \frac{\ln(N_{\boldsymbol{p}})}{A_{\boldsymbol{p}}} + \frac{A_{\boldsymbol{s}}t}{\ln(N_{\boldsymbol{s}})} \\ m_2 &= A_{\boldsymbol{p}}/(A_{\boldsymbol{s}}t + A_{\boldsymbol{p}}) \\ t &= \frac{\max(I_{i\in\boldsymbol{s}} - \min(I_{i\in\boldsymbol{s}}))}{(\sum I_{i\in\boldsymbol{n}})/N_{\boldsymbol{n}}} \\ R^{12} &= m_1/m_2. \end{aligned} \tag{3.1}$$

It is calculated from a baseline-corrected Raman spectrum on the basis of three spectral regions: peak region ($\boldsymbol{p}$), silent region ($\boldsymbol{s}$), and region used for normalization ($\boldsymbol{n}$). The number of data points contained in these three regions is termed $N_{\boldsymbol{p}}$, $N_{\boldsymbol{s}}$, and $N_{\boldsymbol{n}}$, respectively. The variables $A_{\boldsymbol{p}}$ and $A_{\boldsymbol{s}}$ denote the area of the peak and silent region, respectively. Term I represents the Raman intensity.

The idea of the definition is that a good baseline correction should yield the least intensity loss for Raman bands (peak regions) and the least fluorescence residuals for silent regions. By this definition, R^{12} is supposed to be smaller for a better baseline correction. The optimal baseline correction is obtained when R^{12} reaches the minimum.

To verify the definition of R^{12}, a grid search procedure was employed to go through three mathematical baseline correction methods and their parameters. Details of the grid search are available in [A1]. The involved three mathematical baseline correction methods are: sensitive nonlinear iterative peak (SNIP) clipping [174, 175], asymmetric least squares

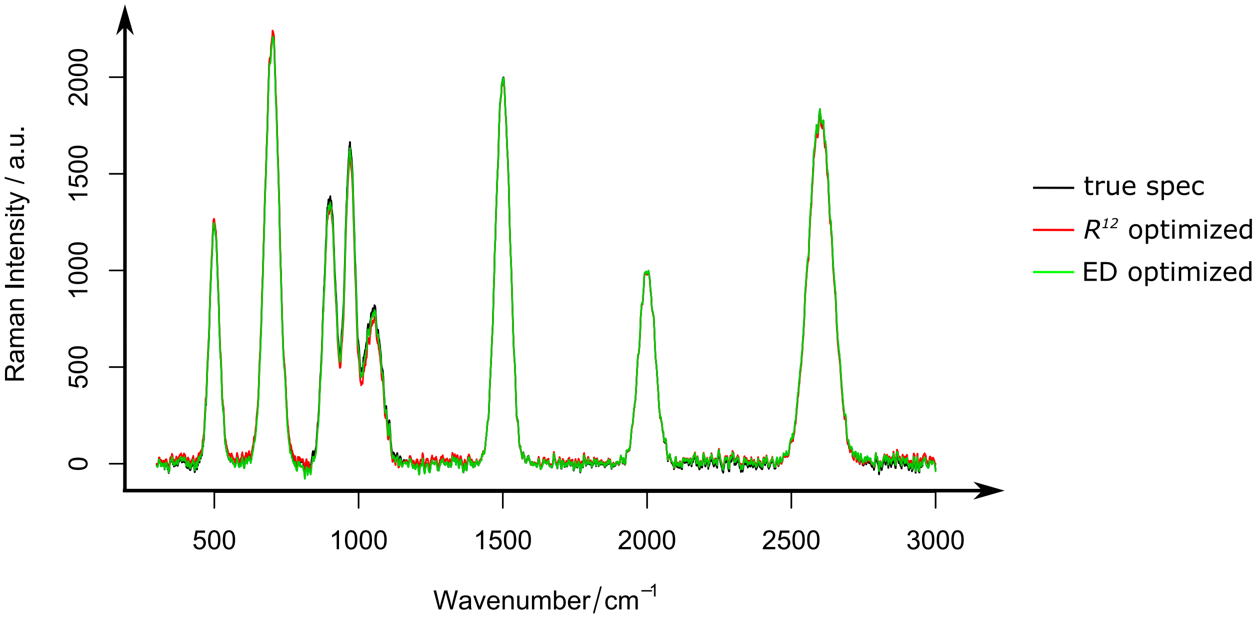


Figure 3.1: Baseline correction optimized by the marker R^{12} and the Euclidean distance (ED) between true and corrected Raman spectra: The optimal baseline correction in the latter case is considered to be the 'gold standard'. The result of the R^{12}-based optimization is consistent to this 'gold standard', demonstrating that R^{12} is a valid figure-of-merit for baseline correction.

(ALS) [176, 177], and modified polynomial (Modpoly) fitting [107]. Baseline correction was performed using each combination of the method and its parameters over the grid search.

As the first step, the validity of R^{12} as a figure-of-merit for baseline correction was verified with an artificial dataset. The artificial dataset was composed of three spectra constructed by the same Raman spectrum and three different fluorescence baselines. Performing the grid search yielded a series of baseline corrections. The optimal baseline correction was selected according to two schemes: (1) the minimal R^{12} and (2) the minimal Euclidean distance (ED) between the true and baseline-corrected Raman spectra. The second scheme was believed to produce the 'gold standard' baseline correction. The results of one of the artificial Raman spectra are shown in Figure 3.1, similar results were observed for the other two spectra. Apparently, the baseline correction featuring the minimal R^{12} was almost the same as the 'gold standard'. This proves that R^{12} is a valid figure-of-merit for baseline correction.

In the next step, the marker R^{12} was employed to optimize the baseline correction of a real-world dataset. The dataset was composed of 1553 Raman spectra measured from three cell types: breast carcinoma-derived tumor cells (MCF-7, BT-20) and acute myeloid leukemia cells (OCI-AML3). Each cell type was measured in nine replicates.

The results from the grid search are shown in Figure 3.2. The bottom plot shows the values of R^{12} calculated from the baseline-corrected spectra for each step of the grid search. The upper two plots visualize the corresponding mean sensitivity of a three-group classifi-

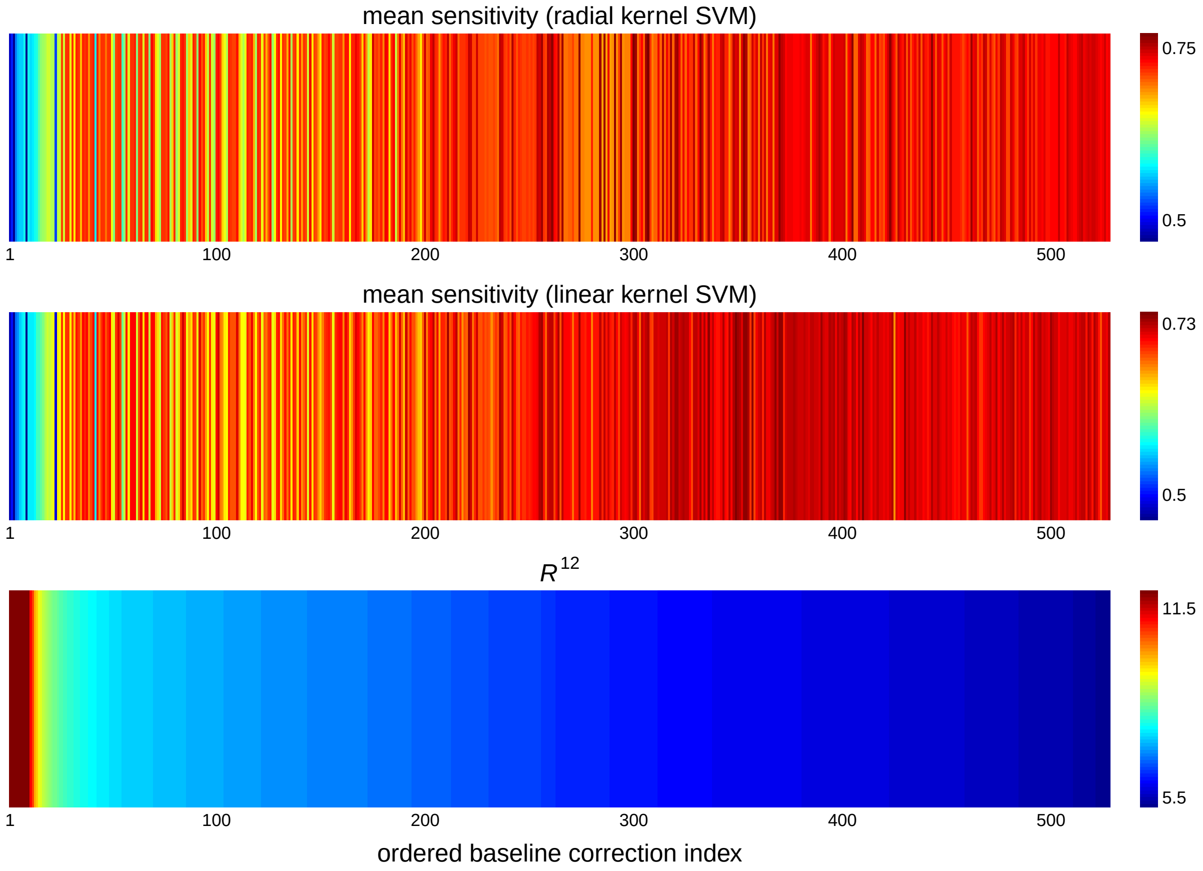


Figure 3.2: Results of R^{12} and classification from the grid search for a three-group task: The value of R^{12} and the mean sensitivity of the three-group classification are calculated after each baseline correction over the grid search. It is evident that the classification performance improves when the R^{12} decreases. The mean sensitivity corresponding to the minimal R^{12} is comparable to the highest values for both linear and kernel SVMs.

cation. The classification was carried out by a combination of principal component analysis (PCA) and support vector machine (SVM) with a linear or radial kernel. It is evident that the mean sensitivity increases as R^{12} decreases. The minimal R^{12} corresponds to the mean sensitivities of 72.7% for linear-kernel SVM and and 73.1% for radial-kernel SVM, which is comparable to the respective highest values of 74.3% and 76.8%. That is to say, a baseline correction optimized by minimizing R^{12} can lead to a classification comparable to the best case.

The estimated baselines corresponding to the minimal R^{12} and the best classification are plotted in Figure 3.3. The estimated baselines leading to the best classification are visibly less reasonable than the baseline with the minimal R^{12}, because the estimated baseline adapted too much to Raman peaks in the former case. That means, the baseline correction selected according to the best classification (i.e., a classification-based optimization) is not necessarily reasonable. In fact, an unreasonable baseline correction can introduce artifacts

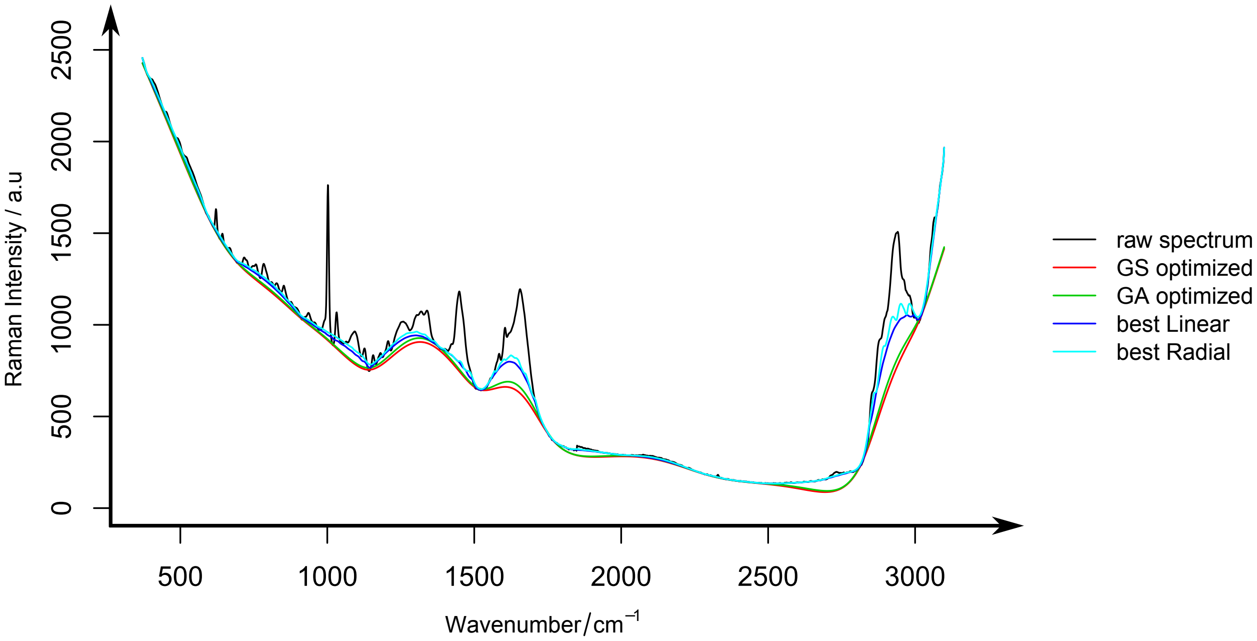


Figure 3.3: Results of estimated baselines corresponding to the minimal R^{12} and the best classification: The R^{12}-based optimization is conducted under the framework of grid search (GS) or genetic algorithm (GA). The two frameworks provide similar results. It is also visible that the R^{12}-based optimization provides more reasonable baseline estimation than the classification-based optimization.

that help the classification. This is the risk of optimizing baseline correction using model-based optimization methods.

The aforementioned grid search can be replaced with a more advanced searching strategy for a faster calculation. A genetic algorithm was utilized in this thesis, in which the combinations of a baseline correction method and the parameters were represented by the chromosomes. The estimated baseline after 150 generations, as shown in Figure 3.3, is consistent with the result of the grid search. However, the genetic algorithm is much faster than the grid search, which saves significant computational costs.

In conclusion, the R^{12}-based optimization provides reasonable baseline corrections and ensures a satisfying classification at the same time. Unlike the classification-based optimization, the R^{12}-based optimization is based on a quantitative marker calculated directly from baseline-corrected spectra and does not require statistical models to be built. Therefore, the R^{12}-based optimization is model independent and does not require a large number of training samples. Last but not least, the R^{12}-based optimization is computationally efficient, especially if an advanced search strategy like genetic algorithms is employed.

3.1.2 Shifted-excitation Raman difference spectroscopy (SERDS)

The automatic optimization method described previously achieves a reasonable mathematical baseline correction. However, the mathematical baseline correction may not be applicable if the fluorescence is too intense and masks the Raman signal. An alternative method is shifted-excitation Raman difference spectroscopy (SERDS), in which two Raman spectra are recorded at two slightly different excitation wavelengths [115]. Ideally, these two measured Raman spectra are composed of identical fluorescence but shifted Raman peaks. The fluorescence is thus removed from the difference spectrum of the two recorded Raman spectra. However, the difference spectrum is difficult to interpret and it is required to recover a fluorescence-free Raman spectrum from the two measured Raman spectra. To deal with this issue, an approach based on a non-negative least squares (NNLS) algorithm is proposed in this subsection. The results are published in [A2].

The idea of the NNLS-based reconstruction is represented by the following equation:

$$\begin{pmatrix} \vec{\boldsymbol{s}}^1 & \vec{\boldsymbol{s}}^2 \end{pmatrix}^T = \begin{bmatrix} \mathbf{I}'_r & \mathbf{I}_f \\ dr \cdot \mathbf{I}_r & \vec{\boldsymbol{df}} \cdot \mathbf{I}_f \end{bmatrix} \times \begin{pmatrix} \vec{\boldsymbol{r}} & \vec{\boldsymbol{f}} \end{pmatrix}^T. \tag{3.2}$$

In this equation, the terms $\vec{\boldsymbol{s}}^1$ and $\vec{\boldsymbol{s}}^2$ denote two Raman spectra measured with different excitation wavelengths. The vectors $\vec{\boldsymbol{r}}$ and $\vec{\boldsymbol{f}}$ represent a fluorescence-free Raman spectrum and a fluorescence baseline, respectively. The identity matrices $\mathbf{I}_r$ and $\mathbf{I}_f$ feature a dimension of $N \times N$, provided N data points are measured for $\vec{\boldsymbol{s}}^1$ and $\vec{\boldsymbol{s}}^2$. The term $\mathbf{I}'_r$ represents a shift matrix of dimension $N \times N$. The shift matrix has 1s on a semi-diagonal and 0s elsewhere. The offset from the semi-diagonal to the main diagonal is equal to the shift between $\vec{\boldsymbol{s}}^1$ and $\vec{\boldsymbol{s}}^2$ counted in spectral data points, which is denoted as parameter m. The direction of the offset (upper or lower to the main diagonal) depends on the direction of the shift from $\vec{\boldsymbol{s}}^2$ to $\vec{\boldsymbol{s}}^1$ (right or left). The scalar dr and the vector $\vec{\boldsymbol{df}}$ represent intensity variations of Raman bands and fluorescence between $\vec{\boldsymbol{s}}^1$ and $\vec{\boldsymbol{s}}^2$, respectively. These two variables are included in order to tackle the influence of intensity variations between $\vec{\boldsymbol{s}}^1$ and $\vec{\boldsymbol{s}}^2$ on the reconstruction. They are calculated by:

$$dr = \frac{\sum_{i=1}^{n} s^1_{lmax_i}}{\sum_{i=1}^{n} s^2_{lmax_i}}, df_k = \frac{spline(s^1_{lmin})_k}{spline(s^2_{lmin})_k}. \tag{3.3}$$

In order to deal with the singularity of $\mathbf{I}'_r$ and improve the stability of the reconstruction, the first m diagonal elements in the shift matrix $\mathbf{I}'_r$ are assigned as 1s. The model after such modification is visualized in Figure 3.4. According to this model, both the fluorescence-free Raman spectrum $\vec{\boldsymbol{r}}$ and the fluorescence baseline $\vec{\boldsymbol{f}}$ can be directly calculated from the two measured Raman spectra $\vec{\boldsymbol{s}}^1$ and $\vec{\boldsymbol{s}}^2$ via a non-negative least squares (NNLS) algorithm.

The spectral reconstruction was first verified by three real-world datasets. The raw data is shown in Figure 3.5. The reconstructed results using NNLS-based method are provided

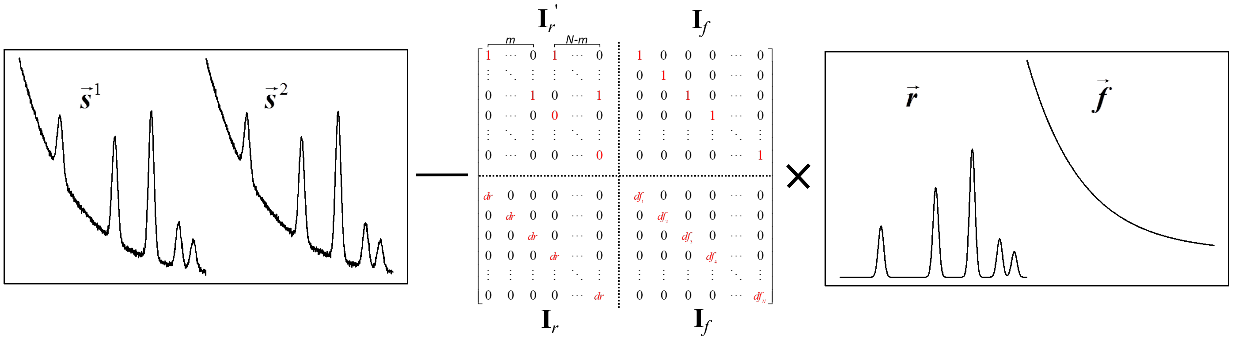


Figure 3.4: Graphic illustration of NNLS-based SERDS spectral reconstruction: A fluorescence-free Raman spectrum and fluorescence baseline can be reconstructed from the two measured Raman spectra via a non-negative least sqaures (NNLS) algorithm.

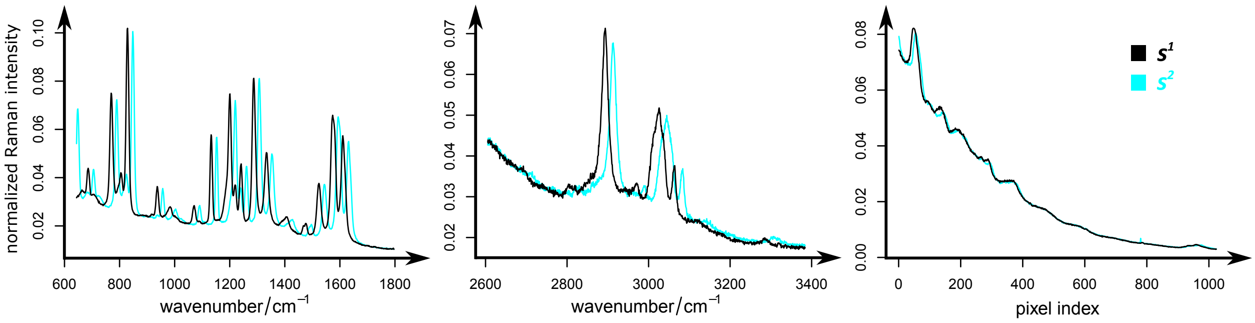


Figure 3.5: The real-world SERDS datasets measured from 4-acetamedophenal (left and middle) and a skin sample from a pig ear (right)

in Figure 3.6. The reconstruction via anti-derivative (antiD) and Fourier transform (FT)-based [98, 118, 120, 121] approaches is displayed as a comparison in Figure 3.7 and Figure 3.8, respectively. Clearly, the spectral resolution was severely degraded by the antiD-based method, because of the implicit average of the antiD-based method. It can also be seen that the FT-based method was corrupted by high-frequency artifacts resulted from the frequency leakage. Both antiD and FT-based reconstructions are overwhelmed by the significant residual fluorescence. In contrast, the NNLS reconstruction is advantageous in terms of the negligible fluorescence-residuals, the unchanged spectral resolution, and the absence of artifacts.

In addition to the real-world datasets, a series of artificial SERDS datasets were employed to quantify the performance of the spectral reconstruction. The artificial datasets were constructed using varying spectral parameters, including the full-width-at-half-maximum (FWHM) of Raman bands, signal-to-noise ratio (SNR), maximal Raman intensity r_{max}, excitation wavelength shift m, and the intensity difference of the fluorescence emission between the two spectra. The simulation was composed of two parts: one without noise and the other with noise. The performance of the reconstruction was quantified from four aspects: the pre-

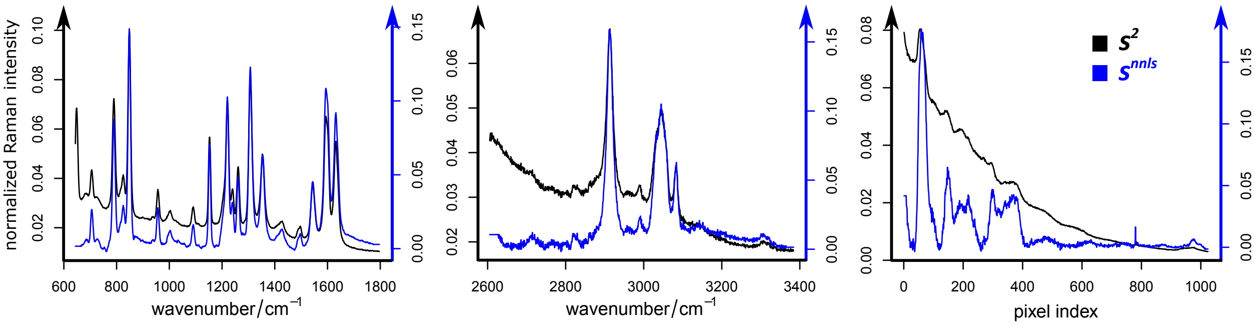


Figure 3.6: Reconstruction results on real-world datasets with NNLS-based method: The reconstructed Raman spectra show negligible fluorescence-residuals and no artifacts. The spectral resolution is almost unchanged after the reconstruction.

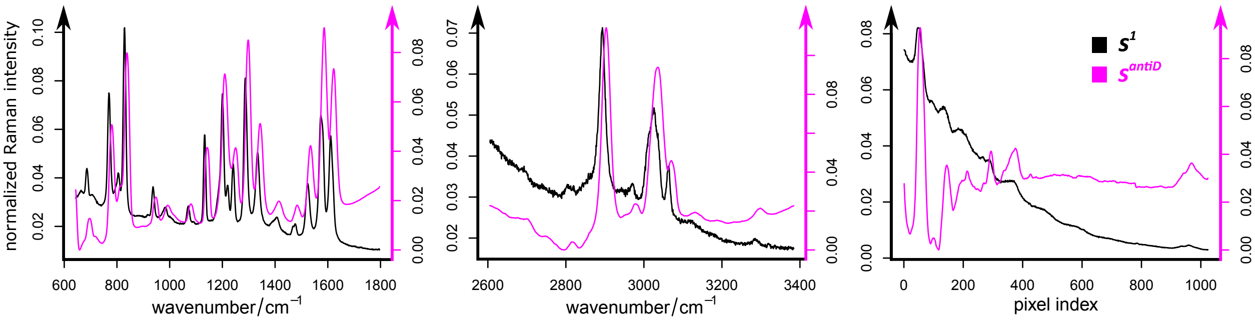


Figure 3.7: Reconstruction results on real-world datasets with antiD-based method: The spectral resolution is evidently degraded after the reconstruction, whereas the fluorescence-residuals are still visible.

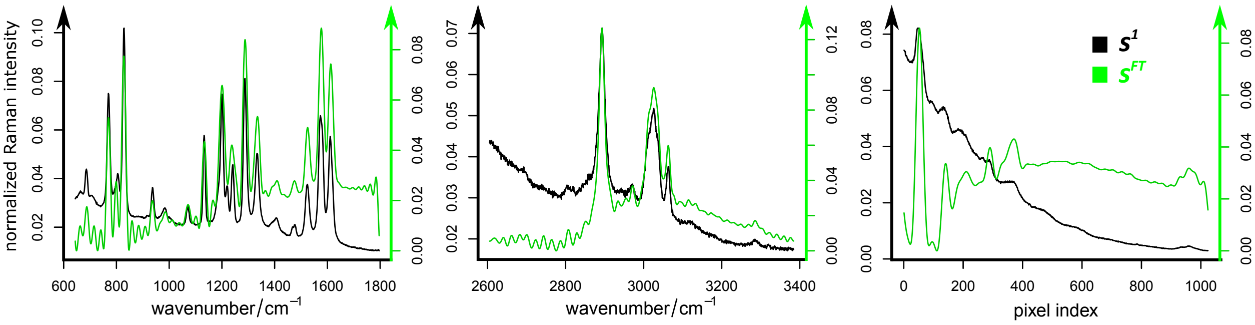


Figure 3.8: Reconstruction results on real-world datasets with FT-based method: The reconstructed spectra are hampered by significant high-frequency artifacts and the severe fluorescence-residuals.

cision of reconstruction, the spectral resolution, the SNR, and the fluorescence-residual.

Details of the simulation and results can be found in [A2]. To briefly summarize, the NNLS-based method can reconstruct the Raman peaks well and with high precision. The reconstructed spectral resolution is not significantly changed compared to the true values. The SNR can be improved after the reconstruction, especially for extremely noisy SERDS datasets. In addition, the NNLS-based reconstruction was almost tolerant to intensity variations between the two measured Raman spectra. Nonetheless, the spectral parameters of the SERDS data are significant factors for the spectral reconstruction. The foremost request is that the excitation wavelength shift m has to match the spectral parameter FWHM. In particular, the NNLS-based approach is proven suitable for datasets with a small FWHM. The antiD is superior to the NNLS-based reconstruction in the case of a larger FWHM, provided the spectral resolution is acceptable after the antiD-based reconstruction.

3.2 Statistical Modeling

After pre-processing, the corrected Raman spectra are used for statistical modeling. A statistical model is built in order to relate the spectral signals to the response variables of interest, such as the disease levels. The statistical model can be one of the following three types: clustering, classification, and regression. The study in this section was focused on classification models. The related work was composed of two parts. In the first part, the framework of a two-layer cross-validation (CV) was established to achieve model optimization and model evaluation. In the second part, a guideline for the application of CV for model optimization and model evaluation was proposed. The investigation was based on 1553 single-cell Raman spectra measured from breast carcinoma-derived tumor cells (MCF-7, BT-20) and acute myeloid leukemia cells (OCI-AML3). Each cell type contains nine technical replicates. The results are published in [A3].

3.2.1 Model construction and validation

The first part of the investigation was conducted on the basis of two binary classification tasks: MCF-7 against BT-20 and MCF-7 against OCI-AML3. The classification model was composed of a dimension reduction using either principal component analysis (PCA) or partial least squares (PLS) and a classifier using either linear discriminant analysis (LDA) or support vector machine (SVM, linear kernel). The classification was performed under the framework of a two-layer cross-validation (CV). As illustrated in Figure 3.9 (left), the two-layer CV was composed of internal and external CV. The internal CV was used to optimize the model parameter (i.e., the number of principal components for PCA (nPC) and the number of latent variables for PLS (nLV)). The external CV was performed for model evaluation. To start CV, the dataset was split into nine folds, each replicate as one fold. For each iteration of the external CV loop, a different fold was taken out as the testing data. The remaining folds were fed into the internal CV loop, where each fold was used once as validation data and predicted by the statistical model built with the remaining folds. The results of the external CV are shown in Figure 3.10. The displayed p values resulted from a paired Wilcoxon test, which was carried out to compare the performance of different models. Only those p values below 0.05 are shown in the plot.

As is shown in Figure 3.10, the four models (PCA-LDA, PCA-SVM, PLS-LDA, and PLS-SVM) performed differently for the task of MCF-7 against OCI-AML3 but almost equally for the task of MCF-7 against BT-20. That is to say, the performance of a model is task dependent. No one single statistical model is superior to the other for all tasks, and model selection (optimization) is always required for each specific application. For the task of MCF-7 against OCI-AML3, PCA-SVM is inferior to PCA-LDA, while PLS-SVM and PLS-LDA are comparable. This means that the employed dimension reduction method is a significant

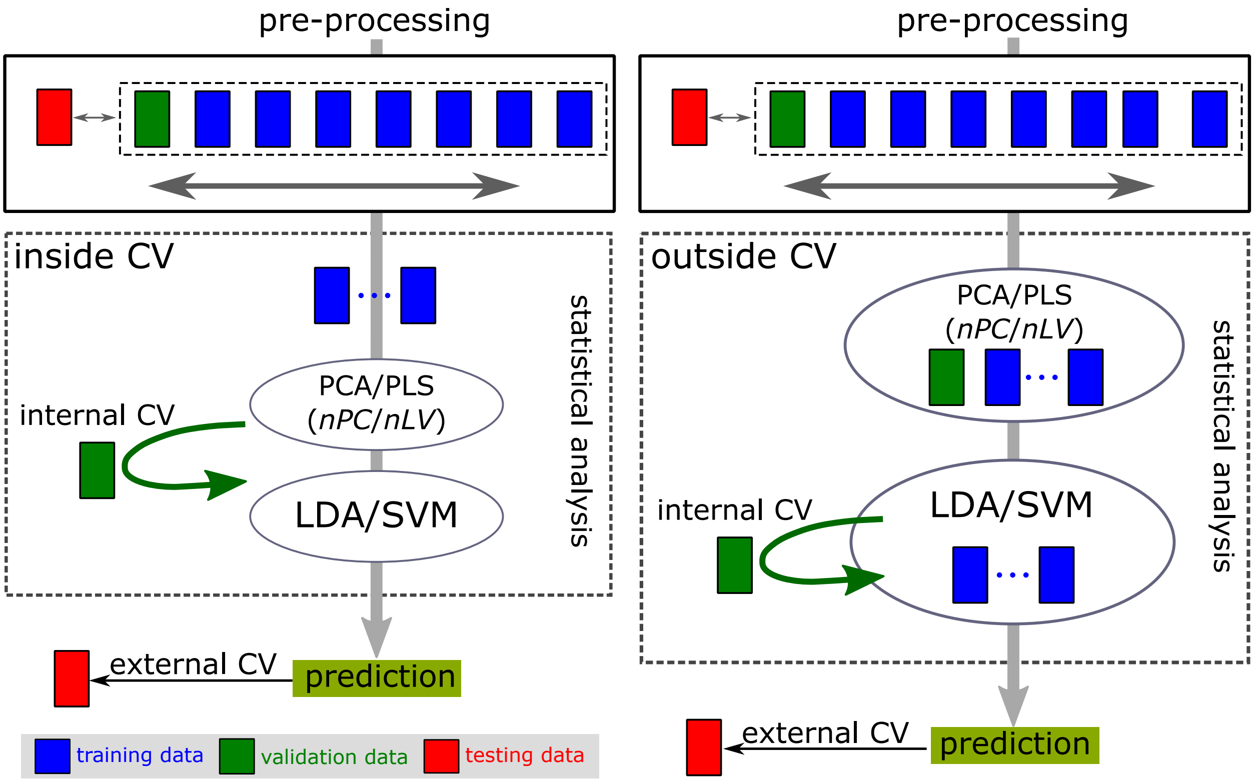


Figure 3.9: Framework of statistical modeling using two-layer cross-validation (CV) for model optimization and evaluation: The statistical model is composed of a dimension reduction (PCA or PLS) and a classifier (LDA or SVM). The two-layer CV is composed of external and internal CV. To conduct CV, the whole dataset is split into multiple folds. For each iteration of the external CV loop, a different fold is taken out as the testing data. The remaining folds are fed into the internal CV loop, where each fold is used once as validation data and predicted by the statistical model built with the remaining folds. The internal CV can be conducted in two manners: inside CV (left panel) and outside CV (right panel). For inside CV, the dimension reduction is carried out excluding the validation data. For outside CV, both training and validation data are employed during the dimension reduction.

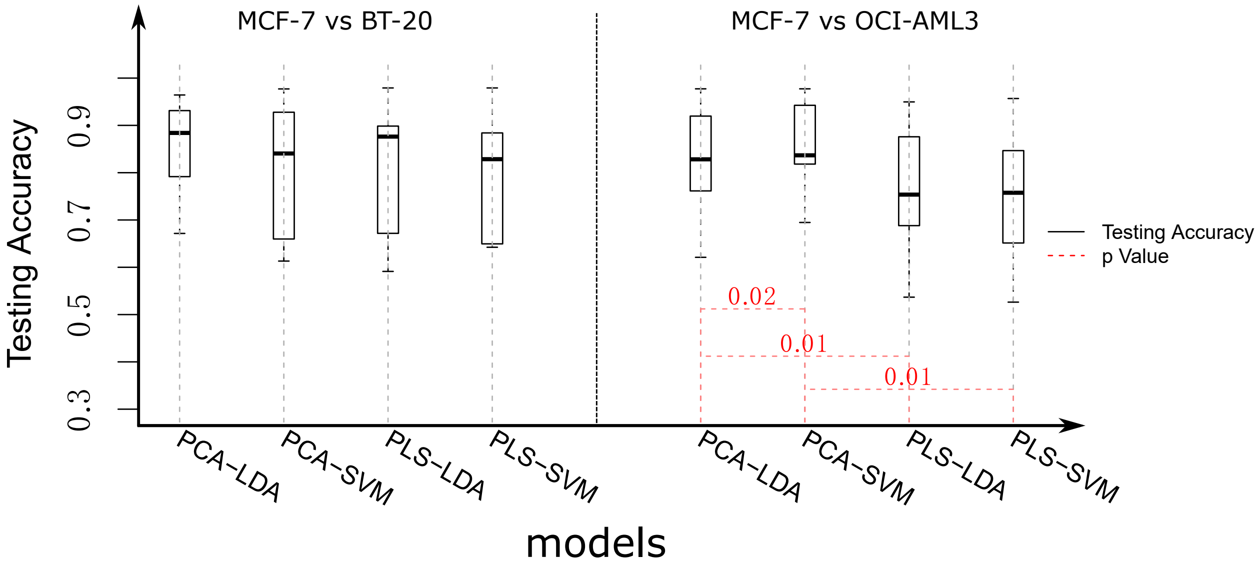


Figure 3.10: Results of testing accuracy from external CV for the two binary tasks. The classification was conducted with different dimension reduction methods (PCA or PLS) and classifiers (SVM or LDA). The model parameter, nPC or nLV, was optimized by internal CV. The displayed p values resulted from a paired Wilcoxon test, which was performed to compare the performance of different models. Only those p values below 0.05 are shown in the plot.

factor for the performance of a classifier. Hence, the model evaluation should be conducted for the dimension reduction and the classifier together. This fact can be further revealed by the investigation in the following subsection.

3.2.2 Common mistakes in cross-validation

Despite the wide application of CV in statistical modeling, it is very common in chemometrics that CV is performed with mistakes [99,100]. The mistakes are mainly manifested from two aspects: improper data splitting and a wrong position of dimension reduction relative to the CV loop. In order to unravel the influence of these two aspects, the internal CV was performed in different cases: k-fold CV, k-replicate CV, inside CV, and outside CV. The k-fold and k-replicate CV corresponded to two different schemes of data split. In the k-fold case, the dataset was split randomly and evenly into k folds. In the k-replicate case, each replicate was used as a different fold. The inside and outside CV represented different positions of the dimension reduction relative to the internal CV loop. As shown in Figure 3.9, the dimension reduction was carried out excluding the validation data for inside CV. In contrast, both training and validation data were employed during the dimension reduction for outside CV. Regardless of the different cases of internal CV, external CV was carried out by taking out each replicate once as testing data. This ensured that the testing data

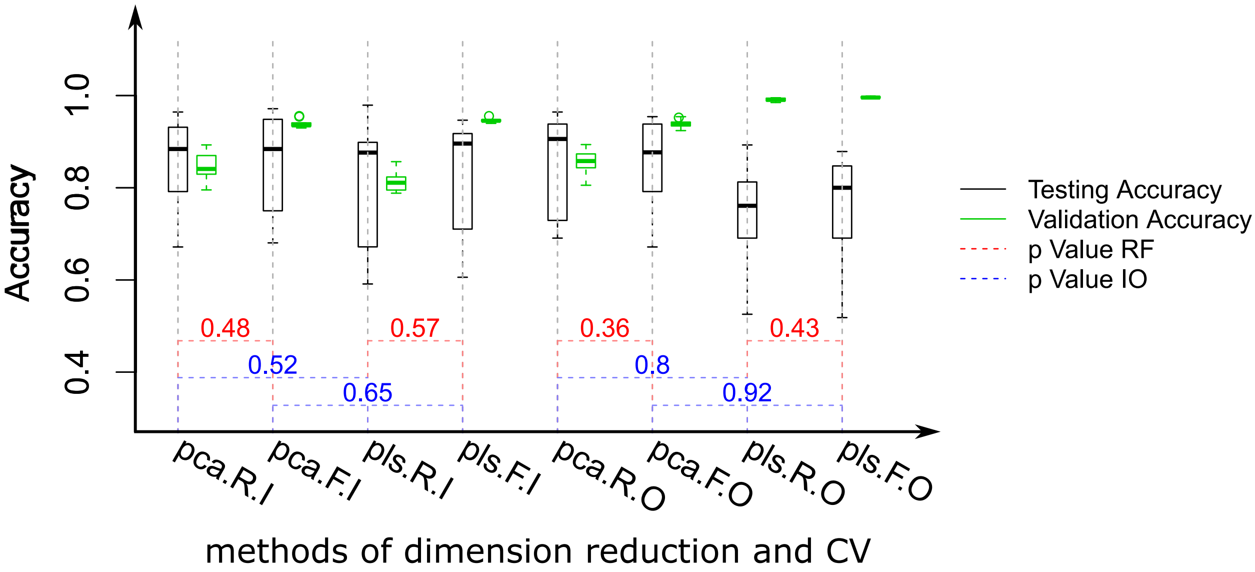


Figure 3.11: Results of the testing accuracies and the validation accuracies in different cases of internal CV: The classification is performed for the task of MCF-7 against BT-20 using LDA combined with PCA or PLS. The labels at the x-axis represent different cases of the internal CV. 'R' and 'F' represent the k-replicate CV and k-fold CV, respectively; while 'I' and 'O' denote the inside CV and outside CV, respectively. The p values are computed by a Wilcoxon test comparing testing accuracies of the models optimized by different internal CVs. The Wilcoxon test was carried out for k-fold CV against k-replicate CV (R-F) and inside CV against outside CV (I-O).

was independent on the training and validation data, hence the testing accuracy reliably benchmarked the performance of the model. The dimension reduction was performed by PCA or PLS, while the classification was conducted using LDA or SVM. The validation accuracy from the internal CV and the testing accuracy from the external CV are plotted in Figure 3.11. Only the results from the task of MCF-7 against BT-20 using LDA are visualized. The results of the other task and SVM are similar and can be found in [A3].

The influence of the two afore-mentioned aspects was unraveled according to the validity of internal CV in terms of model evaluation and model optimization. The validity of internal CV in model evaluation was verified by comparing the validation accuracy to the testing accuracy. As shown in Figure 3.11, the validation accuracy was consistent to the testing accuracy in the case of the k-replicate CV combined with the inside CV ('pca.R.I' and 'pls.R.I'). That is to say the model was reliably evaluated by the internal CV in this case. In all other cases, the validation accuracy was higher than the testing accuracy, meaning the model was over-estimated by the internal CV. The over-estimation was more severe for PLS, a supervised method. Therefore, the dimension reduction must be included inside the internal CV loop, especially for supervised dimension reduction methods. To state differently, the dimension reduction and classifier have to be evaluated together, which is consistent with the conclusion in the previous subsection. The other important factor for reliable model evaluation is that the data split has to be done considering the hierarchy of samples. Data from the same replicate should be treated as one data package (i.e., k-replicate CV).

The validity of internal CV in model optimization was verified by comparing the testing accuracies of the models optimized by different internal CVs (k-fold CV, k-replicate CV, inside CV, and outside CV). To do so, a Wilcoxon test was conducted for each pair of the internal CVs (i.e., k-fold against k-replicate CV and inside against outside CV). The p values are shown in Figure 3.11. The 'p value R-F' corresponds to the comparison between the k-fold and k-replicate CV, while the 'p value I-O' relates to the comparison between the inside and outside CV. The p values are all above 0.05, indicating that the models optimized by the different internal CVs performed almost equally. To state more straightforwardly, the data split scheme and the position of the dimension reduction are less influential if the (internal) CV is used for model optimization compared to model evaulation.

3.3 Model Transfer

The previous sections focused on the issues in chemometrics including pre-processing, statistical modeling, and reliable model evaluation. This section will focus on the model transfer technique. It aims to tackle failed prediction of a constructed model for new data if the new data is severely different from the training data. Such failed prediction is often observed in Raman spectroscopy-based biological investigations, in which spectral differences between new and training data are usually comparable to or larger than spectral variations caused by biological changes of interest. These undesirable spectral differences originate from inter-individual or inter-replicate variations and cannot be removed by the spectrometer calibration. It is infeasible to build a new model on the new data because of the requirement of a large amount of new training samples, especially if new training samples are inaccessible. Instead, model transfer can be employed to enable the trained model to successfully predict the new data [101,168].

In the context of model transfer, training data and new data is denoted as primary and secondary data, respectively. A model trained with primary data is called primary model. The capability of a primary model to predict secondary data is termed model transferability. The proposed model transfer approaches are outlined in subsection 3.3.1. The verification of these approaches is described in subsection 3.3.2. The related work can be found in publications [A4-A6].

The investigation described in this section was based on a Raman spectral dataset measured on four devices from bacterial spores belonging to three species (*B. mycoides*, *B. subtilis*, and *B. thuringienses*). All Raman spectra were calibrated by a spectrometer calibration and were baseline-corrected. The wavenumber and intensity calibration were conducted using the standard material 4-acetamidophenol and SRM 2242, respectively [28,122,123]. With this dataset, the model transfer approaches were applied to transfer the trained model from device to device. However, they can also be employed to transfer the trained model from patient to patient or from replicate to replicate. Such application was demonstrated by the work of [B3] (see chapter 4) [55], in which the model transfer was performed between different biological replicates.

3.3.1 Model transfer approaches

The model transfer approaches proposed herein can be divided into two categories: data-based and model-based model transfer. Data-based approaches aim to eliminate undesired spectral variations between primary and secondary data so that the secondary data can be predicted by the primary model as well. Model-based approaches aim to build a primary model that is robust to the undesired spectral differences between primary and secondary data.

3.3.1.1 Data-based model transfer

As a straightforward way of model transfer, data-based model transfer works by eliminating the spectral differences between primary and secondary data. One of the examples is replicate extended multiplicative signal correction (replicate EMSC) [109, 178, 179]. In this method, the term 'replicate' can refer to data measured from different replicates/individuals (patients), or with different devices. Herein a replicate refers to the dataset measured on the same device, unless otherwise stated.

The model of replicate EMSC is shown by:

$$I_\nu = a + b \cdot m_\nu + d_1 \cdot \nu + d_2 \cdot \nu^2 + \cdots + d_n \cdot \nu^n + \sum_{k=1}^{N} g_k \cdot p_{k\nu} + e_\nu. \tag{3.4}$$

This equation models a spectrum I_ν around a reference spectrum (m_ν) and represents the residuals with a constant offset (a), polynomial profiles ($d_i \cdot \nu^i$), and $g_k \cdot p_{k\nu}$. The reference spectrum I_ν is usually the mean spectrum of the dataset. The polynomials ($d_i \cdot \nu^i$) are used to fit the fluorescence baseline within the Raman spectrum. Term $g_k \cdot p_{k\nu}$ represents the spectral variations over different replicates, which is calculated as the first N loadings of PCA constructed on the mean spectra of different replicates. The coefficients a, b, $\boldsymbol{d}(d_1, d_2, \cdots, d_n)$, and $\boldsymbol{g}(g_1, g_2, \cdots, g_N)$ are fitted for each spectrum by a least squares algorithm and the corrected spectrum is obtained via:

$$I_\nu^c = (I_\nu - a - d_1 \cdot \nu - d_2 \cdot \nu^2 - \cdots - d_n \cdot \nu^n - \sum_{k=1}^{N} g_k \cdot p_{k\nu})/b. \tag{3.5}$$

In this way, the inter-replicate spectral variations are eliminated from the corrected Raman spectra.

It is particularly noteworthy that the polynomial terms in Eq. (3.4) were omitted in this thesis since the Raman spectra were already baseline corrected. The reference spectrum m_ν was calculated as the mean spectrum of all replicates. To calculate the inter-replicate spectral variations ($p_{k\nu}$), the mean Raman spectrum was computed for each replicate and collected into one matrix. A PCA was performed on this matrix following column-wise mean centering. The first N loadings were used as $p_{k\nu}$ of Eq. (3.4).

3.3.1.2 Model-based model transfer

In addition to data-based methods, model transfer can also be achieved with model-based approaches. One such approach is based on Tikhonov regularization (TR) [172], in which the training (primary) dataset ($\mathbf{X}$, $\mathbf{y}$) is augmented with a few secondary samples ($\mathbf{L}$, $\mathbf{y}^*$) (transfer spectra). The augmentation can be done in two ways: TR$_1$ in Eq. (3.6) and TR$_2$ in Eq. (3.7).

$$\begin{pmatrix} \mathbf{y} \\ \lambda \mathbf{y}^* \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \lambda \mathbf{L} \end{pmatrix} \boldsymbol{b}. \tag{3.6}$$

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \\ \lambda \mathbf{y}^{*} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \eta \mathbf{I} \\ \lambda \mathbf{L} \end{pmatrix} \boldsymbol{b}. \tag{3.7}$$

In these equations, parameter λ is introduced to balance the sample sizes of primary and transfer datasets. Variable η and identity matrix $\mathbf{I}$ in TR_2 are used to tackle the singularity problem of TR_1 caused by the possible linearity between primary and transfer spectra. In the study of this thesis, the matrix $\mathbf{L}$ was composed of six transfer spectra randomly selected from the secondary dataset, every two belonging to one group. The matrix $\mathbf{X}$ was composed of primary data, which consists of the Raman spectra from the devices other than the secondary device. In addition, fifteen spectra were randomly selected from the secondary data as optimization spectra, every five belonging to one group. Variables λ and η were optimized by maximizing the prediction of the optimization spectra.

The TR-based methods require the group information of transfer and optimization data. Therefore, they are categorized as supervised model transfer methods. Such model transfer cannot be applied if the group information of secondary data is not accessible. Instead, an unsupervised model transfer is needed, such as the methods based on spectral augmentation (SA) and score movement of principal component analysis (MS) described in the following.

The idea of SA-based model transfer is to enlarge the training space by enforcing wavenumber shifts and intensity variations into training data. The amount of wavenumber shifts (sh_ν) is formulated by $sh_\nu = a_1 + a_2\nu + a_3\nu^2 + a_4\nu^3 + a_5\nu^4$ ($a_i \in [-1, 1]$), where the coefficients ($\boldsymbol{a}$) are generated randomly. The intensity modification function $\boldsymbol{R}$ is a polynomial $R_\nu = c_1 + c_2\nu + c_3\nu^2 + c_4\nu^3$ fitted from 20 randomly selected wavenumbers $\boldsymbol{\nu}^{smp}$ ($min(\boldsymbol{\nu}) < \nu_i^{smp} < max(\boldsymbol{\nu})$, $i = 1, 2, 3, \cdots, 20$) and 20 random values $\boldsymbol{b}$ ($b_i \in [0.5, 2]$, $i = 1, 2, 3, \cdots, 20$). $\boldsymbol{sh}$ and $\boldsymbol{R}$ are recalculated using newly generated $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\nu}^{smp}$ for the modification of each primary spectrum. The primary model is trained with modified primary spectra.

MS-based model transfer corrects undesired spectral variations between primary and secondary data in the score space of PCA. A PCA model is first constructed with primary dataset and used to predict secondary dataset. The scores of primary dataset are moved according to Eq. (3.8), where $\mathbf{X}$, $\overline{\boldsymbol{X}}$, $\mathbf{T}$, and $\mathbf{V}$ represent the spectral matrix, mean spectrum, scores, and loadings, respectively. The superscripts pr, sc, and trs stand for the primary, secondary, and transfer datasets, respectively. The prefix m means the average spectrum of the corresponding dataset.

$$\begin{aligned} \mathbf{X}^{pr} &= \mathbf{T}^{pr}\mathbf{V}^{T}, \mathbf{T}^{sc} = \mathbf{X}^{sc}\mathbf{V} \\ \boldsymbol{T}^{mpr} &= \overline{\boldsymbol{X}}^{pr}\mathbf{V}, \boldsymbol{T}^{mtrs} = \overline{\boldsymbol{X}}^{trs}\mathbf{V} \\ \mathbf{T}^{pr} &= \mathbf{T}^{pr} - (\boldsymbol{T}^{mpr} - \boldsymbol{T}^{mtrs}). \end{aligned} \tag{3.8}$$

3.3.2 Validation of model transfer approaches

To verify the proposed model transfer approaches, a three-group classification was performed based on a partial least squares regression (PLSR) to separate the three bacterial species. In addition, a classifier composed of PCA and SVM (PCA-SVM) was constructed to verify the MS and SA methods. The classification was carried out with a leave-one-device-out cross-validation. That is, each device was used once as the secondary dataset and predicted by the model built with the other three devices. The results of the prediction are visualized in Figure 3.12. The first two columns show the prediction without model transfer, where the classifiers PCA-SVM and PLSR were employed, respectively. The other columns correspond to the prediction using different model transfer methods, as labeled along the x-axis.

As demonstrated in Figure 3.12, the spectrometer calibration provided better transferability than the case without spectrometer calibration, but the improvement was limited. With addition model transfer by the proposed methods, the model transferability was further enhanced by a large scale. The prediction of the first device was generally worse than the prediction of the other devices. The reason for this is that the samples measured on this device were cultivated on a different substrate, leading to severe spectral differences in the data measured on the first and the other three devices. However, the prediction for the first device was still satisfying using the TR-based method. The performance of replicate EMSC was better if more loadings were involved in the replicate EMSC model (3.4).

The model transfer methods were further verified by comparing the prediction after model transfer (i.e., primary prediction) with the prediction by a model built directly on the secondary data (i.e., secondary prediction). To do so, the data from the first device was employed to train the model. The secondary prediction was obtained by a leave-one-batch-out cross-validation and visualized in the first two columns in Figure 3.13, corresponding to the models of PCA-SVM and PLSR, respectively. The solid line represents the mean sensitivity over the three species, while the gray shade denotes the maximum and minimum of the three sensitivities. The dashed line marks to the best results of the secondary prediction. The other columns in Figure 3.13 show the results of the primary prediction using different model transfer methods, i.e., the model was built on the other three devices and used to predict the first device. It is evident that the primary prediction is comparable or even superior to the best secondary prediction when replicate EMSC and TR-based model transfer methods are employed. The primary prediction via the MS-based method is less adequate in this case, due to the large spectral variations between the Raman spectra from the first (secondary) and other devices.

Overall, all the model transfer approaches can improve the model transferability. TR-based methods offer adequate prediction but require the group information of the secondary dataset. Thus, they are inapplicable if the group information of the secondary data is unknown. On the contrary, the unsupervised methods (SA, MS, and replicate EMSC) offer very promising results and do not require the group information of the secondary data. These unsupervised approaches make it possible to transfer the primary model to unknown patients, which is very useful in Raman spectroscopy-based biological diagnostics.

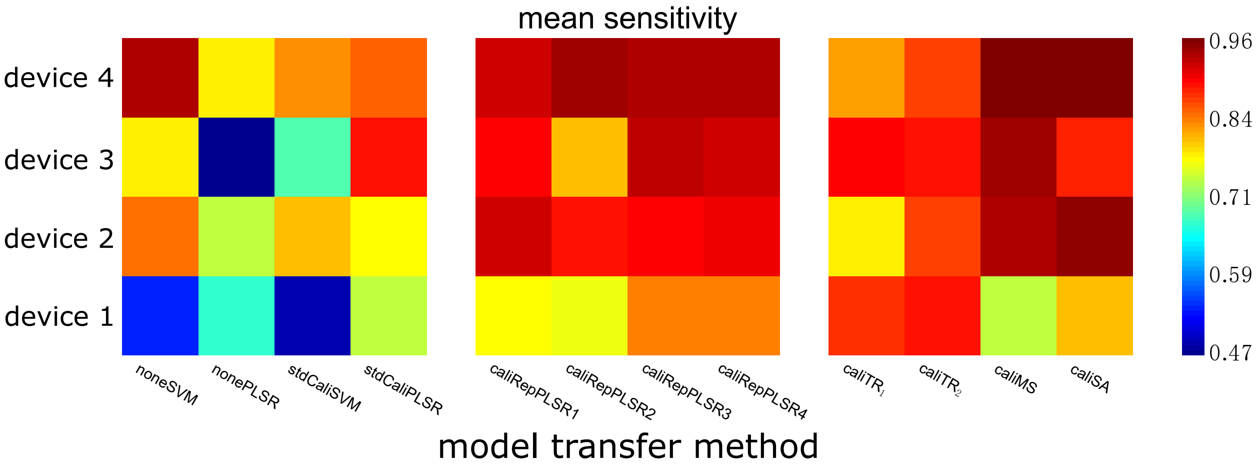


Figure 3.12: Prediction of the secondary dataset in different cases of model transfer: Each device is predicted once by the model built with the other three devices. The prediction is improved by all model transfer methods. The replicate EMSC performs better if more loadings are involved in the replicate EMSC model.

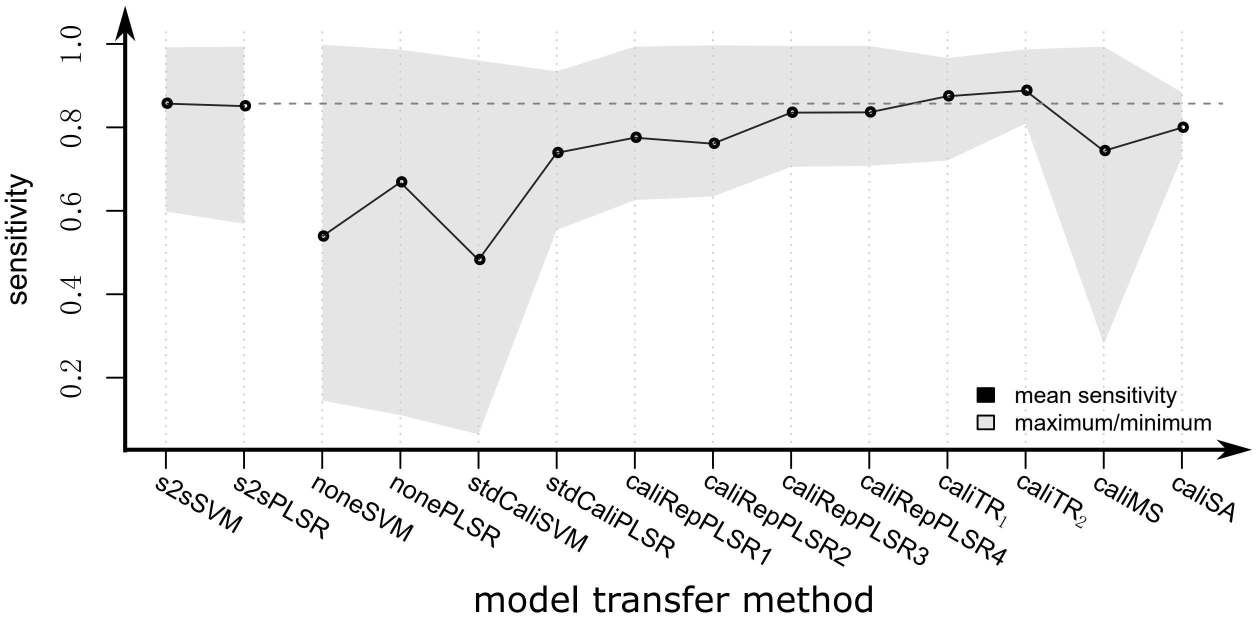


Figure 3.13: Comparison between the results of the primary and secondary prediction for the first device: The first two columns show the secondary prediction obtained by a leave-one-batch-out cross-validation on the first device. The other columns represent primary prediction of the first device by the model built based on the other three devices when different model transfer approaches are employed.

4. Additional Work

In addition to the work being discussed in chapter 3, publications from other investigations during PhD study are listed in the following (ordered by year of publishment):

[B1] S. Guo, S. Pfeifenbring, T. Meyer, G. Ernst, F. von Eggeling, V. Maio, D. Massi, R. Cicchi, F. S. Pavone, J. Popp and T. Bocklitz
Multimodal image analysis in tissue diagnostics for skin melanoma
Journal of Chemometrics, 2018, 32, e2963.

[B2] O. Chernavskaia, S. Guo, T. Meyer, N. Vogler, D. Akimov, S. Heuke, R. Heintzmann, T. Bocklitz and J. Popp
Correction of mosaicking artefacts in multimodal images caused by uneven illumination
Journal of Chemometrics, 2017, 31, e2901.

[B3] V. Kumar B.N., S. Guo, T. Bocklitz, P. Rösch and J. Popp
Demonstration of carbon catabolite repression in naphthalene degrading soil bacteria via Raman spectroscopy based stable isotope probing
Analytical Chemistry, 2016, 88, 7574-7582.

[B4] T. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler, and J. Popp
Raman based molecular imaging and analytics: a magic bullet for biomedical applications!?
Analytical Chemistry, 2016, 88, 133-151.

5. Summary

As mentioned in the chapter 1, chemometrics has become an essential tool in Raman spectroscopy-based biological investigations and significantly enhanced the sensitivity of Raman spectroscopy-based detection. However, there are some open issues on applying chemometrics in Raman spectroscopy-based biological investigations. An automatic procedure is needed to optimize the parameters of the mathematical baseline correction. Spectral reconstruction algorithm is required to recover a fluorescence-free Raman spectrum from the two Raman spectra measured with different excitation wavelengths for the shifted-excitation Raman difference spectroscopy (SERDS) technique. Guidelines are necessary for reliable model optimization and rigorous model evaluation to ensure high accuracy and robustness in Raman spectroscopy-based biological detection. Computational methods are required to enable a trained model to successfully predict new data that is significantly different from the training data due to inter-replicate variations. These tasks were tackled in this thesis. The related investigations were related to three main topics: baseline correction, statistical modeling, and model transfer.

Baseline Correction

Baseline correction refers to the removal of fluorescence emission in measured Raman spectra. The related investigations were conducted from two aspects: mathematical baseline correction and experimental baseline correction.

For mathematical baseline correction, the fluorescence baseline is estimated mathematically and subtracted from the acquired Raman spectrum. In this type of baseline correction, parameters of the methods have to be carefully optimized for each specific dataset. Improper baseline correction can degrade the performance of subsequent analysis. A manual parameter selection is possible but time-consuming and subjective. Therefore, an automatic optimization procedure was proposed, which was based on a quantitative marker R^{12} defined as the figure-of-merit of baseline correction. The marker R^{12} was first verified by three artificial Raman spectra, with which the R^{12}-optimized baseline correction was compared to the 'gold standard'. The 'gold standard' referred to the baseline correction providing the smallest Euclidean distance between the true and baseline-corrected Raman spectrum.

The R^{12}-optimized baseline correction was shown to be consistent with the ‘gold standard’, demonstrating that the marker R^{12} is a valid figure-of-merit for baseline correction. In addition, the R^{12} was employed to optimize the baseline correction for a real-world dataset measured from three cell types (MCF-7, BT-20, and OCI-AML3). The R^{12}-based baseline correction optimization was benchmarked by a three-group classification. According to the results, the R^{12}-based optimization led to a reasonable baseline correction, as well as a satisfying classification performance. Since the marker R^{12} is calculated directly from baseline-corrected Raman spectra, R^{12}-based optimization does not require to build a classification model. Therefore, the optimization is model independent and does not need a large amount of training data. Last but not least, the R^{12}-based optimization is computationally efficient, especially if a fast-searching strategy such as genetic algorithms is used.

For experimental baseline correction, the fluorescence baseline is removed via instrumental modification. SERDS is one of the examples, in which two Raman spectra are recorded at two slightly different excitation wavelengths and the fluorescence is removed from the difference between these two Raman spectra. The difference spectrum is difficult to interpret and it is necessary to reconstruct a fluorescence-free Raman spectrum from the two measured Raman spectra. Existing spectral reconstruction approaches, which are based on anti-derivative and Fourier transform, suffer from drawbacks such as fluorescence residual, spectral resolution loss, and high-frequency artifacts. Therefore, a new spectral reconstruction algorithm was developed based on non-negative least squares (NNLS). According to the results of the three real-world datasets, the NNLS-based method was shown to provide fluorescence-free spectral reconstruction without significantly losing spectral resolution or introducing artifacts. In addition, the performance of the NNLS-based reconstruction was quantified on the basis of artificial datasets. The quantification included four aspects: the precision of reconstruction, spectral resolution, signal-to-noise ratio (SNR), and residual fluorescence. The artificial datasets were constructed with varied SNR, full-width-half-maximum (FWHM), maximal Raman intensity, excitation wavelength shift, and fluorescence variations between the two spectra. It was demonstrated that the NNLS-based method can recover Raman peaks with high precision, unchanged spectral resolution, and improved SNR. The NNLS-based reconstruction was almost tolerant of intensity variations between the two measured Raman spectra. Moreover, it was proven that the excitation wavelength shift has to match the spectral parameter FWHM in order to ensure good spectral reconstruction.

Statistical Modeling

Statistical modeling means to translate Raman spectral signal into high-level information like disease level. The investigation in this thesis was based on Raman spectra measured from three cell types, with which two binary classifications were constructed: MCF-7 against

BT-20 and MCF-7 against OCI-AML3. The statistical model was composed of dimension reduction and classification. Dimension reduction was conducted with either principal component analysis (PCA) or partial least squares (PLS), belonging to supervised and unsupervised methods, respectively. The classification was carried out with either linear discriminant analysis (LDA) or support vector machine (SVM). A two-layer cross-validation (CV) was established, in which the internal CV was used to optimize the model parameters (nPC/nLV) and the external CV was used for evaluating the optimized model. It was demonstrated that the performance of a statistical model is data dependent. No one single statistical model is always superior to the other. In addition, the employed dimension reduction method is a significant factor for the performance of a classifier. The model evaluation should be done for both dimension reduction and the classifier together. Subsequently, a guideline of reliable model optimization and evaluation was proposed. It is related to two significant factors of applying CV in statistical modeling: the data splitting scheme and the position of dimension reduction relative to the CV loop. In the case of model evaluation, the dimension reduction has to be done inside the CV loop, especially if a supervised dimension reduction method is used. More importantly, the data split has to be performed at the highest hierarchy of the sampling. Data from the same biological replicate should be treated as one data package. Both aspects are less influential in the case of model optimization.

Model Transfer

The aim of model transfer is to enable the primary model to successfully predict new (secondary) data that is significantly different from training (primary) data. This is extremely necessary in Raman spectroscopy-based biological applications, where new data often bears significant spectral changes compared to training data due to inter-replicate/individual variations. The model transfer approaches developed in this thesis were based on two mechanisms: data-based model transfer and model-based model transfer. In the former case, undesirable spectral variations between secondary and primary data were estimated and removed. The related approach is replicate extended multiplicative signal correction (replicate EMSC). In the latter case, statistical models resilient to undesirable spectral differences were constructed, as was done by Tikhonov regularization (TR), score movement (MS), and spectral augmentation (SA) methods. In particular, the approaches based on MS, SA, and replicate EMSC belong to unsupervised model transfer and do not require the response information of new data. They have especially great potential in biological diagnostics where the label information of new patients is to be predicted and unknown. The verification of these model transfer methods was based on the classification of Raman spectra from three bacterial spore species (*B. mycoides*, *B. subtilis*, and *B. thuringienses*) measured on four devices. The prediction was obtained via leave-one-device-out cross-validation (i.e., data from each device was used once as secondary data and predicted with a model built based on the other three devices). As was shown, all of the proposed model transfer methods could significantly improve the prediction on secondary data comparing to that without model transfer. Subsequently, the model transfer approaches were verified by comparing the primary prediction with the secondary prediction. The primary prediction refers to predict data of one device by a model built with the other devices. The secondary prediction means to predict data of one device with a model built on this device. Accordingly, the primary predictions using TR and replicate EMSC-based model transfer were comparable or even superior to the secondary prediction.

Above all, this thesis is a step further in resolving the open issues of chemometrics in Raman spectroscopy-based biological investigations. The benefits of the work are four-fold: (1) Automatic parameter optimization helps to effectively handle a massive amount of data and build a ‘one-key’ system for Raman spectroscopy-based biological diagnosis without human intervention. (2) Spectral reconstruction in SERDS technique makes it more convenient to investigate biological samples featuring extremely intense fluorescence emission. (3) Guidelines of model evaluation can help to build robust statistical models and hence to reduce the risk of false diagnosis and improve the reliability of medical diagnosis. (4) Model transfer enables statistical models to predict new data measured from a different individual/replicate, which largely reduces the cost and time required to measure new training samples. In the meantime, model transfer makes it easier to handle data measured by different laboratories. All these benefits are important and highly useful to push Raman-related techniques into the clinical environment.

6. Zusammenfassung

Wie im Kapitel 1 erwähnt, ist die Chemometrie zu einem essentiellen Werkzeug für biologische Untersuchungen mittels der Raman-Spektroskopie geworden und hat die Sensitivität der Raman-spektroskopischen Detektion erheblich verbessert. Es gibt jedoch einige offene Fragen, welche die Anwendung der Chemometrie in Raman-spektroskopischen Untersuchungen biologischer Proben betreffen. Zum Beispiel wird eine automatische Prozedur benötigt, um die Parameter einer mathematischen Basislinienkorrektur zu optimieren. Ein SERDS-Rekonstruktionsalgorithmus ist erforderlich, um ein Fluoreszenz-freies Raman-Spektrum aus den zwei Raman-Spektren zu extrahieren, welche bei der *Shifted-excitation*-Raman-Differenz-Spektroskopie (SERDS) gemessen werden. Des Weiteren sind Richtlinien erforderlich, welche eine zuverlässige Modelloptimierung und eine rigorose Modellevaluation erlauben. Durch diese Richtlinien wird eine hohe Genauigkeit und Robustheit der Raman-spektroskopischen Detektion biologischer Proben gewährleistet. Computergestützte Methoden sind nötig, um mit einem trainierten Modell erfolgreich neue Daten, die sich aufgrund von Inter-Replikat-Variationen signifikant von den Trainingsdaten unterscheiden, vorherzusagen. Diese vier Probleme sind Beispiele für offene Fragen in der Chemometrie und diese vier Probleme wurden in dieser Arbeit behandelt. Die damit verbundenen Untersuchungen bezogen sich auf drei Hauptthemen: die Basislinienkorrektur, die statistische Modellierung und der Modelltransfer.

Basislinienkorrektur

Die Basislinienkorrektur wird eingesetzt, um den Fluoreszenz-Untergrund aus gemessenen Raman-Spektren zu entfernen. Die damit verbundenen Untersuchungen wurden hinsichtlich zweier Korrekturtypen durchgeführt: den mathematischen Basislinienkorrekturen und den experimentellen Basislinienkorrekturen.

Bei der mathematischen Basislinienkorrektur wird die Basislinie mathematisch geschätzt und von den gemessenen Raman-Spektren subtrahiert. Bei dieser Art der Basislinienkorrektur müssen die Parameter der Methoden für jeden Datensatz sorgfältig angepasst und optimiert werden. Werden nicht adäquate Paramater gewählt, resultiert eine ungeeignete Basislinienkorrektur, welche die Leistung der nachfolgenden Analyse beeinträchtigen kann.

Eine manuelle Parameterauswahl ist möglich, jedoch zeitaufwendig und subjektiv. Daher wurde in dieser Arbeit ein automatisches Optimierungsverfahren erforscht, welches auf einem quantitativen Marker R^{12} basiert. Dieser Marker ist definiert, um als Qualitätsmarker der Basislinienkorrektur verwendet zu werden. Der Marker R^{12} wurde zuerst durch drei künstliche Raman-Spektren getestet, indem die mittels R^{12}-optimierte Basislinienkorrektur mit dem ‚Goldstandard‘ verglichen wurde. Der ‚Goldstandard‘ wurde als die Basislinienkorrektur definiert, welche den kleinsten euklidischen Abstand zwischen dem wahren und dem basislinienkorrigierten Raman-Spektrum lieferte. Die R^{12}-optimierte Basislinie erwies sich als konsistent mit dem ‚Goldstandard‘, was zeigt, dass der Marker R^{12} ein gutes Qualitätsmaß für eine Basislinienkorrektur ist. Zusätzlich wurde der Marker R^{12} verwendet, um die Basislinienkorrektur für einen realen Datensatz, der aus Raman-Spektren von drei Zelltypen (MCF-7, BT-20 und OCI-AML3) bestand, zu optimieren. Die auf R^{12}-basierende Basislinienkorrekturoptimierung wurde durch eine Drei-Gruppen-Klassifizierung evaluiert. Den Ergebnissen zufolge, resultierte die R^{12}-basierende Optimierung in einer Basislinienkorrektur, welche zu einer zufriedenstellenden Klassifikationsleistung führte. Da der Marker R^{12} direkt aus den Basislinien-korrigierten Raman-Spektren berechnet wird, erfordert die R^{12}-basierende Optimierung keine Konstruktion eines Klassifikationsmodells. Daher ist die Optimierung modellunabhängig und benötigt keinen großen Trainingsdatensatz. Nicht zuletzt ist die R^{12}-basierende Optimierung rechnerisch effizient, insbesondere wenn eine schnelle Suchstrategie wie ein genetischer Algorithmen verwendet wird.

Bei einer experimentellen Basislinienkorrektur wird die Fluoreszenzbasislinie durch instrumentelle Modifikation des Messgeräts entfernt. SERDS ist eine experimentelle Basislinienkorrektur bei der zwei Raman-Spektren mit zwei leicht unterschiedlichen Anregungswellenlängen aufgenommen werden. Die Fluoreszenz wird dann durch Differenzbildung zwischen diesen beiden Raman-Spektren entfernt. Das Differenzspektrum ist aber schwer zu interpretieren. Daher muss aus den beiden gemessenen Raman-Spektren ein fluoreszenzfreies Raman-Spektrum rekonstruiert werden. Vorhandene SERDS-Rekonstruktionsverfahren, wie die Integration und die Fourier-Transformations-basierende Rekonstruktion, besitzen Nachteile, wie zum Beispiel das die Fluoreszenz nicht vollständig korrigiert werden kann, das die spektrale Auflösung sich verschlechtert und das es zum sogenannten *Frequency-Leakage* kommt. Daher wurde ein neuer SERDS-Rekonstruktionsalgorithmus basierend auf dem *Non-Negative-Least-Square*-Algorithmus (NNLS) entwickelt. Basierend auf den Ergebnissen für drei reale Datensätze konnte gezeigt werden, dass die NNLS-basierende Methode eine fluoreszenzfreie Rekonstruktion ermöglicht, ohne dass die spektrale Auflösung signifikant abnimmt oder offensichtliche Artefakte auftreten. Zusätzlich wurde die Leistungsfähigkeit der NNLS-basierenden Rekonstruktion anhand von künstlichen Datensätzen quantifiziert. Die Quantifizierung umfasste vier Aspekte: die Genauigkeit der Rekonstruktion, die spektrale Auflösung, das Signal-zu-Rausch-Verhältnis (SNR) und die Restfluoreszenz. Die künstlichen

Datensätze wurden mit unterschiedlichem SNR, verschiedenem *Full-width-half-maximum* (FWHM), unterschiedlicher maximalen Raman-Intensität, verschiedener Anregungswellenlängenverschiebung und verschiedener Fluoreszenzvariation zwischen den beiden Raman-Spektren konstruiert. Es konnte gezeigt werden, dass das NNLS-basierende Verfahren Raman-Banden mit hoher Präzision, unveränderter spektraler Auflösung und verbessertem SNR wiederherstellen kann. Die NNLS-basierende Rekonstruktion war stabil gegenüber Intensitätsschwankungen zwischen den beiden gemessenen Raman-Spektren. Darüber hinaus wurde nachgewiesen, dass die Anregungswellenlängenverschiebung mit dem spektralen Parameter FWHM übereinstimmen muss, um eine gute spektrale Rekonstruktion zu gewährleisten.

Statistische Modellierung

Die statistische Modellierung wird benötigt, um die Raman-Spektren in *High-Level*-Informationen, wie das Krankheitsniveau oder den Zelltyp, zu übersetzen. Die Untersuchungen in dieser Arbeit basierten auf Raman-Spektren, welche an drei Zelltypen gemessen wurden. Mit diesen Raman-Spektren wurden zwei binäre Klassifikationssysteme konstruiert: MCF-7 gegen BT-20 und MCF-7 gegen OCI-AML3. Dabei bestand das statistische Modell aus Dimensionsreduktion und einem Klassifikationsmodell. Die Dimensionsreduktion wurde entweder mit einer Hauptkomponentenanalyse (PCA) oder einer *Partial-Least-Square*-Regression (PLS) durchgeführt, die jeweils zu den überwachten und nicht überwachten Dimensionsreduktionsmethoden gehören. Die Klassifizierung wurde entweder mit einer linearen Diskriminanzanalyse (LDA) oder mit einer *Support Vector Machine* (SVM) durchgeführt. Es wurde eine zweistufige Kreuzvalidierung (CV) etabliert, in der die interne CV zur Optimierung der Modellparameter (nPC/nLV) und die externe CV zur Evaluation des optimierten Modells verwendet wurde. Es konnte gezeigt werden, dass die Qualität eines statistischen Modells datenabhängig ist. Kein statistisches Modell ist dem anderen Modell immer überlegen. Darüber hinaus ist das verwendete Dimensionsreduktionsverfahren ein wesentlicher Faktor für die Leistungsfähigkeit eines Klassifikators. Es konnte auch gezeigt werden, dass die Modellevaluation immer für die Dimensionsreduktion und den Klassifikator zusammen durchgeführt werden sollte. Die Resultate dieser Untersuchungen wurden abschließend zu einer Leitlinie für eine zuverlässige Modelloptimierung und -evaluation zusammengefasst. Dabei sind die wichtigsten Faktoren, welche die Anwendung einer CV bei der statistischen Modellierung beeinflussen, das Datenaufteilungsschema und die Position der Dimensionsreduktion relativ zur CV-Schleife. Bei der Modellevaluation muss die Dimensionsreduktion innerhalb der CV-Schleife erfolgen, insbesondere wenn eine überwachte Dimensionsreduktionsmethode verwendet wird. Noch wichtiger ist, dass die Datenaufteilung auf der höchsten Stufe der Proben-Hierarchie durchgeführt werden muss. Daten aus demselben biologischen Replikat oder demselben Patient sollten als ein Datenpaket behandelt werden. Bei der Modelloptimierung haben das Datenaufteilungsschema und die Position der Dimensionsreduktion relativ zur CV-Schleife einen kleineren Einfluss.

Modelltransfer

Das Ziel des Modelltransfers besteht darin, das Primärmodell in die Lage zu versetzen, neue Sekundärdaten, welche sich signifikant von den Trainingsdaten (Primärdaten) unterscheiden, erfolgreich vorherzusagen. Dies ist bei biologischen Anwendungen der Raman-Spektroskopie extrem wichtig, da neue Daten häufig signifikante spektrale Änderungen im Vergleich zu den Trainingsdaten aufweisen. Diese Änderungen resultieren aus Replikat-Variationen und der immer existenten biologischen Variation. Die in dieser Arbeit entwickelten Modelltransfer-Ansätze basieren auf zwei Mechanismen: dem datenbasierenden Modelltransfer und dem modellbasierenden Modelltransfer. Im ersten Fall werden unerwünschte spektrale Schwankungen zwischen sekundären und primären Daten geschätzt und entfernt. Der in dieser Arbeit verwandte Ansatz ist die Replikat-*Extended-multiplicative-signal-correction* (Replikat-EMSC). Beim modellbasierenden Modelltransfer werden statistische Modelle konstruiert, die gegenüber unerwünschten spektralen Unterschieden stabil sind, wie dies durch die Tikhonov-*regularization* (TR)-, *Score-movement* (MS)- und *Spectral-augmentation* (SA)-Methode durchgeführt wird. Insbesondere gehören die MS-, SA- und Replikat-EMSC-basierenden Methoden, zu den nicht-überwachten Modelltransfer-Methoden und benötigen keine Label-Informationen der sekundären Daten. Damit haben sie ein besonders großes Potenzial um in der biologischen Diagnostik eingesetzt zu werden, da bei diesen Anwendungen die Label-Informationen eines neuen Patienten vorhergesagt werden sollen und somit unbekannt sind. Das Testen dieser Modelltransfermethoden basierte auf der Klassifizierung von Raman-Spektren von drei bakteriellen Sporenarten (*B. mycoides*, *B. subtilis* und *B. thuringienses*), die mittels vier Messgeräten gemessen wurden. Die Vorhersage wurde durch eine *Leave-one-device-out*-Kreuzvalidierung durchgeführt. Dabei werden Daten von jedem Gerät einmal als sekundäre Daten verwendet und mit einem Modell vorhergesagt, das basierend auf den Daten der anderen drei Geräte erstellt wurde. Es konnte gezeigt werde, dass alle vorgeschlagenen Modelltransfermethoden die Vorhersage der sekundären Daten im Vergleich zu der Vorhersage ohne Modelltransfer signifikant verbessern. Anschließend wurden die Modellübertragungsansätze mit einander verglichen, indem die primäre Vorhersage mit der sekundären Vorhersage verglichen wurde. Die primäre Vorhersage bezieht sich auf die Vorhersage von Daten eines Geräts durch ein Modell, das mit den Daten der anderen Geräte erstellt wurde. Die sekundäre Vorhersage bedeutet, dass Daten eines Geräts mit einem Modell vorhergesagt werden, welches mit Daten dieses Geräts erstellt wurde. Es konnte gezeigt werden, dass die primäre Vorhersagen unter Verwendung von TR- und Replikat-EMSC-basierendem Modelltransfer der sekundären Vorhersage vergleichbar oder sogar überlegen war.

Diese Arbeit ist ein weiterer Schritt, um offenen Probleme der Chemometrie in Raman-Spektroskopie-basierenden biologischen Untersuchungen zu lösen. Die in der Arbeit vorgestellte Lösungen haben verschiedene Vorteile: (1) Die automatische Parameteroptimierung für mathematische Basislinienkorrekturen hilft eine große Menge an Daten effektiv zu verarbeiten und ein vollautomatisierte Analysesysteme für die auf der Raman-Spektroskopie basierende biologischen Diagnostik aufzubauen. (2) Die spektrale Rekonstruktion für SERD-Spektren macht es einfacher, biologische Proben mit extremer Fluoreszenz zu untersuchen. (3) Die Leitlinien für die Modellevaluierung können dazu beitragen, robuste statistische Modelle zu erstellen und somit das Risiko falscher Diagnosen zu verringern. Auch kann die Zuverlässigkeit der medizinischen Diagnostik basierend auf der Raman-Spektroskopie verbessert werden. (4) Der Modelltransfer ermöglicht es mit statistischen Modellen neue Daten, die von einem anderen Individuum oder Replikat stammen, genauer vorherzusagen. Dadurch muss kein neues Modell konstruiert werden, was die Kosten- und den Zeitaufwand erheblich reduziert.

Bibliography

[1] AJ Tatem, DJ Rogers, and SI Hay. Global transport networks and infectious disease spread. *Advances in Parasitology*, 62:293–343, 2006.

[2] AH Lichtenstein, LJ Appel, M Brands, M Carnethon, S Daniels, HA Franch, B Franklin, P Kris-Etherton, WS Harris, B Howard, et al. Diet and lifestyle recommendations revision 2006: a scientific statement from the American heart association nutrition committee. *Circulation*, 114(1):82–96, 2006.

[3] FB Hu. Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes Care*, 34(6):1249–1257, 2011.

[4] T Fulop, A Larbi, JM Witkowski, J McElhaney, M Loeb, A Mitnitski, and G Pawelec. Aging, frailty and age-related diseases. *Biogerontology*, 11(5):547–563, 2010.

[5] J Popp, VV Tuchin, A Chiou, and SH Heinemann. *Handbook of Biophotonics*, volume 1. Wiley VCH, 2012.

[6] TF Massoud and SS Gambhir. Molecular imaging in living subjects: seeing fundamental biological processes in a new light. *Genes & Development*, 17(5):545–580, 2003.

[7] T Hussain and QT Nguyen. Molecular imaging for cancer diagnosis and surgery. *Advanced drug Delivery Reviews*, 66:90–100, 2014.

[8] P Padmanabhan, A Kumar, S Kumar, RK Chaudhary, and B Gulyás. Nanoparticles in practice for molecular-imaging applications: an overview. *Acta Biomaterialia*, 41:1–16, 2016.

[9] MA Pysz, SS Gambhir, and JK Willmann. Molecular imaging: current status and emerging strategies. *Clinical Radiology*, 65(7):500–516, 2010.

[10] R Weissleder. A clearer vision for in vivo imaging. *Nature Biotechnology*, 19:316–317, 2001.

[11] J Popp. *Ex-vivo and In-vivo Optical Molecular Pathology.* Wiley, 2014.

[12] JW Lichtman and J Conchello. Fluorescence microscopy. *Nature Methods*, 2(12):910, 2005.

[13] B Huang, H Babcock, and X Zhuang. Breaking the diffraction barrier: super-resolution imaging of cells. *Cell*, 143(7):1047–1058, 2010.

[14] BO Leung and KC Chou. Review of super-resolution fluorescence microscopy for biology. *Applied Spectroscopy*, 65(9):967–980, 2011.

[15] SJ Sahl and WE Moerner. Super-resolution fluorescence imaging with single molecules. *Current Opinion in Structural Biology*, 23(5):778–787, 2013.

[16] PIH Bastiaens and A Squire. Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell. *Trends in Cell Biology*, 9(2):48–52, 1999.

[17] H Wallrabe and A Periasamy. Imaging protein molecules using FRET and FLIM microscopy. *Current Opinion in Biotechnology*, 16(1):19–27, 2005.

[18] W Min, CW Freudiger, S Lu, and XS Xie. Coherent nonlinear optical imaging: beyond fluorescence microscopy. *Annual Review of Physical Chemistry*, 62:507–530, 2011.

[19] J Ries and P Schwille. Fluorescence correlation spectroscopy. *BioEssays*, 34(5):361–368, 2012.

[20] WR Zipfel, RM Williams, R Christie, AY Nikitin, BT Hyman, and WW Webb. Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proceedings of the National Academy of Sciences*, 100(12):7075–7080, 2003.

[21] F Helmchen and W Denk. Deep tissue two-photon microscopy. *Nature Methods*, 2(12):932, 2005.

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

[23] R Pepperkok and J Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7(9):690, 2006.

[24] EA Carter and HG Edwards. Biological applications of Raman spectroscopy. *Infrared and Raman Spectroscopy of Biological Materials*, 24:421, 2001.

[25] CV Raman and KS Krishnan. A change of wave-length in light scattering. *Nature*, 121(3051):619, 1928.

[26] CV Raman and KS Krishnan. A new type of secondary radiation. *Nature*, 121(3048):501–502, 1928.

[27] G Landsberg. Eine neue Erscheinung bei der Lichtzerstreuung in Krystallen. *Naturwissenschaften*, 16:558, 1928.

[28] RL McCreery. *Raman Spectroscopy for Chemical Analysis*, volume 157. John Wiley & Sons, 2000.

[29] M Schmitt and J Popp. Raman spectroscopy at the beginning of the twenty-first century. *Journal of Raman Spectroscopy*, 37(1-3):20–28, 2006.

[30] H Gremlich and B Yan. *Infrared and Raman Spectroscopy of Biological Materials*. CRC Press, 2000.

[31] JT Edsall. Raman spectra of amino acids and related compounds I. the ionization of the carboxyl group. *The Journal of Chemical Physics*, 4(1):1–8, 1936.

[32] FS Parker. *Applications of Infrared, Raman, and Resonance Raman Spectroscopy in Biochemistry*. Springer Science & Business Media, 1983.

[33] RG Gould. The LASER, light amplification by stimulated emission of radiation. In *The Ann Arbor Conference on Optical Pumping, the University of Michigan*, volume 15, page 128, 1959.

[34] PH Lissberger and WL Wilcock. Properties of all-dielectric interference filters. II. filters in parallel beams of light incident obliquely and in convergent beams. *Journal of the Optical Society of America*, 49(2):126–130, 1959.

[35] WS Boyle and GE Smith. Charge coupled semiconductor devices. *Bell Labs Technical Journal*, 49(4):587–593, 1970.

[36] LP Choo-Smith, HGM Edwards, HP Endtz, JM Kros, F Heule, H Barr, JS Robinson, HA Bruining, and GJ Puppels. Medical applications of Raman spectroscopy: from proof of principle to clinical implementation. *Biopolymers*, 67(1):1–9, 2002.

[37] HJ Butler, L Ashton, B Bird, G Cinque, K Curtis, J Dorney, K Esmonde-White, NJ Fullwood, B Gardner, PL Martin-Hirsch, et al. Using Raman spectroscopy to characterize biological materials. *Nature Protocols*, 11(4):664–687, 2016.

[38] JL Koenig. Raman spectroscopy of biological molecules: a review. *Journal of Polymer Science: Macromolecular Reviews*, 6(1):59–177, 1972.

[39] WL Peticolas. Applications of Raman spectroscopy to biological macromolecules. *Biochimie*, 57(4):417–428, 1975.

[40] RC Lord. Strategy and tactics in the Raman spectroscopy of biomolecules. *Applied Spectroscopy*, 31(3):187–194, 1977.

[41] Z Movasaghi, S Rehman, and IU Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007.

[42] T Bocklitz, S Guo, O Ryabchykov, N Vogler, and J Popp. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? *Analytical Chemistry*, 88(1):133–151, 2016.

[43] C Kendall, M Isabelle, F Bazant-Hegemark, J Hutchings, L Orr, J Babrah, R Baker, and N Stone. Vibrational spectroscopy: a clinical tool for cancer diagnostics. *Analyst*, 134(6):1029–1045, 2009.

[44] CA Owen, J Selvakumaran, I Notingher, G Jell, LL Hench, and MM Stevens. In vitro toxicology evaluation of pharmaceuticals using Raman micro-spectroscopy. *Journal of Cellular Biochemistry*, 99(1):178–186, 2006.

[45] I Notingher. Raman spectroscopy cell-based biosensors. *Sensors*, 7(8):1343–1358, 2007.

[46] S Stöckel, S Meisel, M Elschner, P Rösch, and J Popp. Identification of bacillus anthracis via Raman spectroscopy and chemometric approaches. *Analytical Chemistry*, 84(22):9873–9880, 2012.

[47] S Pahlow, S Meisel, D Cialla-May, K Weber, P Rösch, and J Popp. Isolation and identification of bacteria by means of Raman spectroscopy. *Advanced Drug Delivery Reviews*, 89:105–120, 2015.

[48] B Lorenz, C Wichmann, S Stöckel, P Rösch, and J Popp. Cultivation-free Raman spectroscopic investigations of bacteria. *Trends in Microbiology*, 25(5):413–424, 2017.

[49] B Kang, MM Afifi, LA Austin, and MA El-Sayed. Exploiting the nanoparticle plasmon effect: observing drug delivery dynamics in single cells via Raman/fluorescence imaging spectroscopy. *ACS Nano*, 7(8):7420–7427, 2013.

[50] SF El-Mashtoly, D Petersen, HK Yosef, A Mosig, A Reinacher-Schick, C Kötting, and K Gerwert. Label-free imaging of drug distribution and metabolism in colon cancer cells by Raman microscopy. *Analyst*, 139(5):1155–1161, 2014.

[51] X Bi, B Rexer, CL Arteaga, M Guo, and A Mahadevan-Jansen. Evaluating HER2 amplification status and acquired drug resistance in breast cancer cells using Raman spectroscopy. *Journal of Biomedical Optics*, 19(2):025001–025001, 2014.

[52] DI Ellis and R Goodacre. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst*, 131(8):875–885, 2006.

[53] B Berry, J Moretto, T Matthews, J Smelko, and K Wiltberger. Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis. *Biotechnology Progress*, 31(2):566–577, 2015.

[54] S Kumar, N Matange, S Umapathy, and SS Visweswariah. Linking carbon metabolism to carotenoid production in mycobacteria using Raman spectroscopy. *FEMS Microbiology Letters*, 362(3):1–6, 2015.

[55] V Kumar BN, S Guo, T Bocklitz, P Rösch, and J Popp. Demonstration of carbon catabolite repression in naphthalene degrading soil bacteria via Raman spectroscopy based stable isotope probing. *Analytical Chemistry*, 88(15):7574–7582, 2016.

[56] E Mistek, L Halámková, KC Doty, CK Muro, and IK Lednev. Race differentiation by Raman spectroscopy of a bloodstain for forensic purposes. *Analytical Chemistry*, 88(15):7453–7456, 2016.

[57] CAF Penido, MTT Pacheco, IK Lednev, and L Silveira. Raman spectroscopy in forensic analysis: identification of cocaine and other illegal drugs of abuse. *Journal of Raman Spectroscopy*, 47(1):28–38, 2016.

[58] DI Ellis, DP Cowcher, L Ashton, S O'Hagan, and R Goodacre. Illuminating disease and enlightening biomedicine: Raman spectroscopy as a diagnostic tool. *Analyst*, 138(14):3871–3884, 2013.

[59] C Kallaway, LM Almond, H Barr, J Wood, J Hutchings, C Kendall, and N Stone. Advances in the clinical application of Raman spectroscopy for cancer diagnostics. *Photodiagnosis and Photodynamic Therapy*, 10(3):207–219, 2013.

[60] W Wang, J Zhao, M Short, and H Zeng. Real-time in vivo cancer diagnosis using Raman spectroscopy. *Journal of Biophotonics*, 8(7):527–545, 2015.

[61] C Krafft and J Popp. The many facets of Raman spectroscopy for biomedical analysis. *Analytical and Bioanalytical Chemistry*, 407(3):699–717, 2015.

[62] K Kong, C Kendall, N Stone, and I Notingher. Raman spectroscopy for medical diagnostics from in-vitro biofluid assays to in-vivo cancer detection. *Advanced Drug Delivery Reviews*, 89:121–134, 2015.

[63] P Matousek and N Stone. Development of deep subsurface Raman spectroscopy for medical diagnosis and disease monitoring. *Chemical Society Reviews*, 45(7):1794–1802, 2016.

[64] N Stone, C Kendall, J Smith, P Crow, and H Barr. Raman spectroscopy for identification of epithelial cancers. *Faraday Discussions*, 126:141–157, 2004.

[65] IP Santos, PJ Caspers, TC Bakker Schut, R van Doorn, V Noordhoek Hegt, S Koljenović, and GJ Puppels. Raman spectroscopic characterization of melanoma and benign melanocytic lesions suspected of melanoma using high-wavenumber Raman spectroscopy. *Analytical Chemistry*, 88(15):7683–7688, 2016.

[66] R Malini, K Venkatakrishna, J Kurien, K M Pai, L Rao, VB Kartha, and CM Krishna. Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: a Raman spectroscopy study. *Biopolymers*, 81(3):179–193, 2006.

[67] EM Barroso, RWH Smits, TC Bakker Schut, I Ten Hove, JA Hardillo, EB Wolvius, RJ Baatenburg de Jong, S Koljenovic, and GJ Puppels. Discrimination between oral cancer and healthy tissue based on water content determined by Raman spectroscopy. *Analytical Chemistry*, 87(4):2419–2426, 2015.

[68] T Meyer, N Bergner, C Bielecki, C Krafft, D Akimov, BFM Romeike, R Reichart, R Kalff, B Dietzek, and J Popp. Nonlinear microscopy, infrared, and Raman microspectroscopy for brain tumor analysis. *Journal of Biomedical Optics*, 16(2):021113–021113, 2011.

[69] M Jermyn, K Mok, J Mercier, J Desroches, J Pichette, K Saint-Arnaud, L Bernstein, M Guiot, K Petrecca, and F Leblond. Intraoperative brain cancer detection with Raman spectroscopy in humans. *Science Translational Medicine*, 7(274):274ra19–274ra19, 2015.

[70] R Stables, G Clemens, HJ Butler, KM Ashton, A Brodbelt, TP Dawson, LM Fullwood, MD Jenkinson, and MJ Baker. Feature driven classification of Raman spectra for real-time spectral brain tumour diagnosis using sound. *Analyst*, 142(1):98–109, 2017.

[71] N Stone, P Stavroulaki, C Kendall, M Birchall, and H Barr. Raman spectroscopy for early detection of laryngeal malignancy: preliminary results. *The Laryngoscope*, 110(10):1756–1763, 2000.

[72] SK Teh, W Zheng, DP Lau, and Z Huang. Spectroscopic diagnosis of laryngeal carcinoma using near-infrared Raman spectroscopy and random recursive partitioning ensemble techniques. *Analyst*, 134(6):1232–1239, 2009.

[73] AS Haka, Z Volynskaya, JA Gardecki, J Nazemi, R Shenk, N Wang, RR Dasari, M Fitzmaurice, and MS Feld. Diagnosing breast cancer using Raman spectroscopy: prospective analysis. *Journal of Biomedical Optics*, 14(5):054023–054023, 2009.

[74] B Abramczyk H, Brozek-Pluska, J Surmacki, J Jablonska-Gajewicz, and R Kordek. Raman 'optical biopsy' of human breast cancer. *Progress in Biophysics and Molecular Biology*, 108(1):74–81, 2012.

[75] Z Huang, A McWilliams, H Lui, DI McLean, S Lam, and H Zeng. Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. *International Journal of Cancer*, 107(6):1047–1052, 2003.

[76] ND Magee, JR Beattie, C Carland, R Davis, K McManus, I Bradbury, DA Fennell, PW Hamilton, M Ennis, JJ McGarvey, et al. Raman microscopy in the diagnosis and prognosis of surgically resected nonsmall cell lung cancer. *Journal of Biomedical Optics*, 15(2):026015–026015, 2010.

[77] J Horsnell, P Stonelake, J Christie-Brown, G Shetty, J Hutchings, C Kendall, and N Stone. Raman spectroscopy – a new method for the intra-operative assessment of axillary lymph nodes. *Analyst*, 135(12):3042–3047, 2010.

[78] GR Lloyd, LE Orr, J Christie-Brown, K McCarthy, S Rose, M Thomas, and N Stone. Discrimination between benign, primary and secondary malignancies in lymph nodes from the head and neck utilising Raman spectroscopy and multivariate analysis. *Analyst*, 138(14):3900–3908, 2013.

[79] BWD de Jong, TC Bakker Schut, K Maquelin, T van der Kwast, CH Bangma, D Kok, and GJ Puppels. Discrimination between nontumor bladder tissue and tumor by Raman spectroscopy. *Analytical Chemistry*, 78(22):7761–7769, 2006.

[80] ROP Draga, MCM Grimbergen, PLM Vijverberg, CFP van Swol, TGN Jonges, JA Kummer, and JLH Ruud Bosch. In vivo bladder cancer diagnosis by high-volume Raman spectroscopy. *Analytical Chemistry*, 82(14):5993–5999, 2010.

[81] D Lin, S Feng, J Pan, Y Chen, J Lin, G Chen, S Xie, H Zeng, and R Chen. Colorectal cancer detection by gold nanoparticle based surface-enhanced Raman spectroscopy of blood serum and statistical analysis. *Optics Express*, 19(14):13565–13577, 2011.

[82] N Vogler, T Bocklitz, F Subhi Salah, C Schmidt, R Bräuer, T Cui, M Mireskandari, FR Greten, M Schmitt, A Stallmach, et al. Systematic evaluation of the biological

variance within the Raman based colorectal tissue diagnostics. *Journal of Biophotonics*, 9(5):533–541, 2016.

[83] A Pallaoro, MR Hoonejani, GB Braun, CD Meinhart, and M Moskovits. Rapid identification by surface-enhanced Raman spectroscopy of cancer cells at low concentrations flowing in a microfluidic channel. *ACS Nano*, 9(4):4328–4336, 2015.

[84] II Patel, J Trevisan, G Evans, V Llabjani, PL Martin-Hirsch, HF Stringfellow, and FL Martin. High contrast images of uterine tissue derived using Raman microspectroscopy with the empty modelling approach of multivariate curve resolution-alternating least squares. *Analyst*, 136(23):4950–4959, 2011.

[85] C Bielecki, T Bocklitz, M Schmitt, C Krafft, C Marquardt, A Gharbi, T Knösel, A Stallmach, and J Popp. Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells. *Journal of Biomedical Optics*, 17(7):0760301–0760308, 2012.

[86] E Ryzhikova, O Kazakov, L Halamkova, D Celmins, P Malone, E Molho, EA Zimmerman, and IK Lednev. Raman spectroscopy of blood serum for Alzheimer's disease diagnostics: specificity relative to other types of dementia. *Journal of Biophotonics*, 8(7):584–596, 2015.

[87] RG Brereton, J Jansen, J Lopes, F Marini, A Pomerantsev, O Rodionova, JM Roger, B Walczak, and R Tauler. Chemometrics in analytical chemistry-part I: history, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry*, 409(25):5891–5899, 2017.

[88] S Wold. Spline functions, a new tool in data-analysis. *Kemisk Tidskrift*, 84(3):34, 1972.

[89] G Weber. Enumeration of components in complex systems by fluorescence spectrophotometry. *Nature*, 190(4770):27–29, 1961.

[90] BR Kowalski. Chemometrics: mathematics and statistics in chemistry. In *NATO Science Series C*, 1984.

[91] P Rösch, M Harz, M Schmitt, and J Popp. Raman spectroscopic identification of single yeast cells. *Journal of Raman Spectroscopy*, 36(5):377–379, 2005.

[92] U Neugebauer, T Bocklitz, JH Clement, C Krafft, and J Popp. Towards detection and identification of circulating tumour cells using Raman spectroscopy. *Analyst*, 135(12):3178–3182, 2010.

[93] N Bergner, T Bocklitz, BFM Romeike, R Reichart, R Kalff, C Krafft, and J Popp. Identification of primary tumors of brain metastases by Raman imaging and support vector machines. *Chemometrics and Intelligent Laboratory Systems*, 117:224–232, 2012.

[94] S Kloß, P Rösch, W Pfister, M Kiehntopf, and J Popp. Toward culture-free Raman spectroscopic identification of pathogens in ascitic fluid. *Analytical Chemistry*, 87(2):937–943, 2014.

[95] W Richardson, D Wilkinson, L Wu, F Petrigliano, B Dunn, and D Evseenko. Ensemble multivariate analysis to improve identification of articular cartilage disease in noisy Raman spectra. *Journal of Biophotonics*, 8(7):555–566, 2015.

[96] T Bocklitz, A Walter, K Hartmann, P Rösch, and J Popp. How to pre-process Raman spectra for reliable and stable models? *Analytica Chimica Acta*, 704(1):47–56, 2011.

[97] J Engel, J Gerretzen, E Szymańska, JJ Jansen, G Downey, L Blanchet, and LMC Buydens. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, 2013.

[98] MT Gebrekidan, C Knipfer, F Stelzle, J Popp, S Will, and A Braeuer. A shifted-excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference. *Journal of Raman Spectroscopy*, 47(2):198–209, 2016.

[99] M Defernez and EK Kemsley. The use and misuse of chemometrics for treating classification problems. *TrAC Trends in Analytical Chemistry*, 16(4):216–221, 1997.

[100] P Harrington. Multiple versus single set validation of multivariate models to avoid mistakes. *Critical Reviews in Analytical Chemistry*, 48(1):33–46, 2018.

[101] JH Kalivas, GG Siano, E Andries, and HC Goicoechea. Calibration maintenance and transfer using Tikhonov regularization approaches. *Applied Spectroscopy*, 63(7):800–809, 2009.

[102] P Heraud, BR Wood, J Beardall, and D McNaughton. Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalgal cells. *Journal of Chemometrics*, 20(5):193–197, 2006.

[103] NK Afseth, VH Segtnan, and JP Wold. Raman spectra of biological samples: a study of preprocessing methods. *Applied Spectroscopy*, 60(12):1358–1367, 2006.

[104] O Ryabchykov, T Bocklitz, A Ramoji, U Neugebauer, M Foerster, C Kroegel, M Bauer, M Kiehntopf, and J Popp. Automatization of spike correction in Raman spectra of biological samples. *Chemometrics and Intelligent Laboratory Systems*, 155:1–6, 2016.

[105] B Brownfield and JH Kalivas. Consensus outlier detection using sum of ranking differences of common and new outlier measures without tuning parameter selections. *Analytical Chemistry*, 89(9):5087–5094, 2017.

[106] Z Zhang, S Chen, Y Liang, Z Liu, Q Zhang, L Ding, F Ye, and H Zhou. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *Journal of Raman Spectroscopy*, 41(6):659–669, 2010.

[107] CA Lieber and A Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied Spectroscopy*, 57(11):1363–1367, 2003.

[108] S Baek, A Park, Y Ahn, and J Choo. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst*, 140(1):250–257, 2015.

[109] NK Afseth and A Kohler. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 117:92–99, 2012.

[110] MW Meyer, JS Lupoi, and EA Smith. 1064nm dispersive multichannel Raman spectroscopy for the analysis of plant lignin. *Analytica Chimica Acta*, 706(1):164–170, 2011.

[111] F Knorr, ZJ Smith, and S Wachsmann-Hogiu. Development of a time-gated system for Raman spectroscopy of biological samples. *Optics Express*, 18(19):20049–20058, 2010.

[112] J Gerretzen, E Szymańska, J Bart, AN Davies, H van Manen, ER van den Heuvel, JJ Jansen, and LMC Buydens. Boosting model performance and interpretation by entangling preprocessing selection and variable selection. *Analytica Chimica Acta*, 938:44–52, 2016.

[113] J Egermann, T Seeger, and A Leipertz. Application of 266-nm and 355-nm Nd: YAG laser radiation for the investigation of fuel-rich sooting hydrocarbon flames by Raman scattering. *Applied Optics*, 43(29):5564–5574, 2004.

[114] R Adami and J Kiefer. Light-emitting diode based shifted-excitation Raman difference spectroscopy (LED-SERDS). *Analyst*, 138(21):6258–6261, 2013.

[115] AP Shreve, NJ Cherepy, and RA Mathies. Effective rejection of fluorescence interference in Raman spectroscopy using a shifted excitation difference technique. *Applied Spectroscopy*, 46(4):707–711, 1992.

[116] MA da Silva Martins, DG Ribeiro, EAP dos Santos, AA Martin, A Fontes, and H da Silva Martinho. Shifted-excitation Raman difference spectroscopy for in vitro and in vivo biological samples analysis. *Biomedical Optics Express*, 1(2):617–626, 2010.

[117] P Matousek, M Towrie, and AW Parker. Fluorescence background suppression in Raman spectroscopy using combined Kerr gated and shifted excitation Raman difference techniques. *Journal of Raman Spectroscopy*, 33(4):238–242, 2002.

[118] P Matousek, M Towrie, and AW Parker. Simple reconstruction algorithm for shifted excitation Raman difference spectroscopy. *Applied Spectroscopy*, 59(6):848–851, 2005.

[119] J Cooper, MF Abdelkader, and K Wise. Method and apparatus for acquiring Raman spectra without background interferences, October 29 2013. US Patent 8,570,507.

[120] J Zhao, MM Carrabba, and FS Allen. Automated fluorescence rejection using shifted excitation Raman difference spectroscopy. *Applied Spectroscopy*, 56(7):834–845, 2002.

[121] I Osticioli, A Zoppi, and EM Castellucci. Shift-excitation Raman difference spectroscopy – difference deconvolution method for the luminescence background rejection from Raman spectra of solid samples. *Applied Spectroscopy*, 61(8):839–844, 2007.

[122] T Bocklitz, T Dörfer, R Heinke, M Schmitt, and J Popp. Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths. *Spectrochimica*

Acta Part A: Molecular and Biomolecular Spectroscopy, 149:544–549, 2015.

[123] T Dörfer, T Bocklitz, N Tarcea, M Schmitt, and J Popp. Checking and improving calibration of Raman spectra using chemometric approaches. *Zeitschrift für Physikalische Chemie*, 225(6-7):753–764, 2011.

[124] MM Carrabba. *Handbook of Vibrational Spectroscopy*, chapter Wavenumber standards for Raman spectrometry. Wiley Online Library, 2006.

[125] I Guyon and A Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[126] P Geladi and BR Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[127] R Li and X Wang. Dimension reduction of process dynamic trends using independent component analysis. *Computers & Chemical Engineering*, 26(3):467–473, 2002.

[128] X Zhang and R Tauler. Application of multivariate curve resolution alternating least squares (MCR-ALS) to remote sensing hyperspectral imaging. *Analytica Chimica Acta*, 762:25–38, 2013.

[129] H Li, Y Liang, Q Xu, and D Cao. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, 648(1):77–84, 2009.

[130] S Wang, D Li, X Song, Y Wei, and H Li. A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, 38(7):8696–8702, 2011.

[131] H Li, Y Liang, D Cao, and Q Xu. Model-population analysis and its applications in chemical and biological modeling. *TrAC Trends in Analytical Chemistry*, 38:154–162, 2012.

[132] Y Yun, W Wang, B Deng, G Lai, X Liu, D Ren, Y Liang, W Fan, and Q Xu. Using variable combination population analysis for variable selection in multivariate calibration. *Analytica Chimica Acta*, 862:14–23, 2015.

[133] W Cai, Y Li, and X Shao. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 90(2):188–194, 2008.

[134] R Diao and Q Shen. Two new approaches to feature selection with harmony search. In *2010 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–7, 2010.

[135] Z Zhang, TWS Chow, and M Zhao. M-Isomap: orthogonal constrained marginal isomap for nonlinear dimensionality reduction. *IEEE Transactions on Cybernetics*, 43(1):180–191, 2013.

[136] VD Silva and JB Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems*, pages 721–728,

2003.

[137] R Shan, W Cai, and X Shao. Variable selection based on locally linear embedding mapping for near-infrared spectral analysis. *Chemometrics and Intelligent Laboratory Systems*, 131:31–36, 2014.

[138] GE Hinton and RR Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[139] W Wang, Y Huang, Y Wang, and L Wang. Generalized autoencoder: a neural network framework for dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 490–497, 2014.

[140] AK Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[141] C Muehlethaler, G Massonnet, and P Esseiva. The application of chemometrics on infrared and Raman spectra as a tool for the forensic analysis of paints. *Forensic Science International*, 209(1-3):173–182, 2011.

[142] T Tolstik, C Marquardt, C Beleites, C Matthäus, C Bielecki, M Bürger, C Krafft, O Dirsch, U Settmacher, J Popp, and A Stallmach. Classification and prediction of HCC tissues by Raman imaging with identification of fatty acids as potential lipid biomarkers. *Journal of Cancer Research and Clinical Oncology*, 141(3):407–418, 2015.

[143] S Khan, R Ullah, A Khan, A Sohail, N Wahab, M Bilal, and M Ahmed. Random forest-based evaluation of Raman spectroscopy for dengue fever analysis. *Applied Spectroscopy*, 71(9):2111–2117, 2017.

[144] M Gniadecka, PA Philipsen, S Wessel, R Gniadecki, HC Wulf, S Sigurdsson, OF Nielsen, DH Christensen, J Hercogova, K Rossen, et al. Melanoma diagnosis by Raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. *Journal of Investigative Dermatology*, 122(2):443–449, 2004.

[145] R Dong, S Weng, L Yang, and J Liu. Detection and direct readout of drugs in human urine using dynamic surface-enhanced Raman spectroscopy and support vector machines. *Analytical Chemistry*, 87(5):2937–2944, 2015.

[146] K Thangavel and A Pethalakshmi. Dimensionality reduction based on rough set theory: a review. *Applied Soft Computing*, 9(1):1–12, 2009.

[147] A Eslami, EM Qannari, A Kohler, and S Bougeard. Multivariate analysis of multiblock and multigroup data. *Chemometrics and Intelligent Laboratory Systems*, 133:63–69, 2014.

[148] S Saha and A Ekbal. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, 85:15–39, 2013.

[149] K Héberger. Sum of ranking differences compares methods or models fairly. *TrAC Trends in Analytical Chemistry*, 29(1):101–109, 2010.

[150] M Bevilacqua and F Marini. Local classification: locally weighted-partial least squares-discriminant analysis (LW-PLS-DA). *Analytica Chimica Acta*, 838:20–30, 2014.

[151] J Schmidhuber. Deep learning in neural networks: an overview. *Neural Networks*, 61:85–117, 2015.

[152] J Liu, M Osadchy, L Ashton, M Foster, CJ Solomon, and SJ Gibson. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst*, 142(21):4067–4074, 2017.

[153] JH Kalivas and J Palmer. Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. *Journal of Chemometrics*, 28(5):347–357, 2014.

[154] P Refaeilzadeh, L Tang, and H Liu. Cross-validation. In *Encyclopedia of Database Systems*, pages 532–538. Springer, 2009.

[155] RA Fisher. *The Design of Experiments.* Oliver And Boyd, 1937.

[156] P Gy. *Sampling for Analytical Purposes.* John Wiley & Sons, 1998.

[157] GP Quinn and MJ Keough. *Experimental Design and Data Analysis for Biologists.* Cambridge University Press, 2002.

[158] S Mukherjee, P Tamayo, S Rogers, R Rifkin, A Engle, C Campbell, TR Golub, and JP Mesirov. Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology*, 10(2):119–142, 2003.

[159] KK Dobbin and RM Simon. Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics*, 8(1):101–117, 2006.

[160] C Beleites, U Neugebauer, T Bocklitz, C Krafft, and J Popp. Sample size planning for classification models. *Analytica Chimica Acta*, 760:25–33, 2013.

[161] L Petersen, P Minkkinen, and KH Esbensen. Representative sampling for reliable data analysis: theory of sampling. *Chemometrics and Intelligent Laboratory Systems*, 77(1):261–277, 2005.

[162] KH Esbensen and P Geladi. Principles of proper validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*, 24(3-4):168–187, 2010.

[163] R Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (Ijcai)*, pages 1137–1145, 1995.

[164] J Shao and D Tu. *The Jackknife and Bootstrap.* Springer Science & Business Media, 2012.

[165] FR Burden, RG Brereton, and PT Walsh. Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy.

Analyst, 122(10):1015–1022, 1997.

[166] C Soneson, S Gerster, and M Delorenzi. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PloS One*, 9(6):e100335, 2014.

[167] JB Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354, 1983.

[168] J Ottaway and JH Kalivas. Feasibility study for transforming spectral and instrumental artifacts for multivariate calibration maintenance. *Applied Spectroscopy*, 69(3):407–416, 2015.

[169] C Liang, H Yuan, Z Zhao, C Song, and J Wang. A new multivariate calibration model transfer method of near-infrared spectral analysis. *Chemometrics and Intelligent Laboratory Systems*, 153:51–57, 2016.

[170] TG Bloemberg, J Gerretzen, A Lunshof, R Wehrens, and LMC Buydens. Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. *Analytica Chimica Acta*, 781:14–32, 2013.

[171] JH Kalivas, B Brownfield, and BJ Karki. Sample-wise spectral multivariate calibration desensitized to new artifacts relative to the calibration data using a residual penalty. *Journal of Chemometrics*, 31(4), 2017.

[172] JH Kalivas. Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *Journal of Chemometrics*, 26(6):218–230, 2012.

[173] M Blackburn, S Ramos, and B Rohrback. Transfer of calibration for classification problems. InfoMetrix, 2002.

[174] CG Ryan, E Clayton, WL Griffin, SH Sie, and DR Cousens. SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 34(3):396–402, 1988.

[175] M Miroslav. *Peaks: Peaks*, 2012. R Package Version 0.2.

[176] PHC Eilers and HFM Boelens. Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, 1(1):5, 2005.

[177] Kristian HL and Bjørn-Helge M. *baseline: baseline correction of spectra*, 2015. R package version 1.2-1.

[178] H Martens, SW Bruun, I Adt, GD Sockalingum, and A Kohler. Pre-processing in biochemometrics: correction for path-length and temperature effects of water in FTIR bio-spectroscopy by EMSC. *Journal of Chemometrics*, 20(8-10):402–417, 2006.

[179] KH Liland, A Kohler, and NK Afseth. Model-based pre-processing in Raman spectroscopy of biological samples. *Journal of Raman Spectroscopy*, 47(6):643–650, 2016.

7. Publications

7.1 Optimization of Raman-Spectrum Baseline Correction in Biological Application (A1)

S. Guo, T. Bocklitz, and J. Popp, *Analyst*, 2016, 141, 2396-2404.

Der Nachdruck der folgenden Publikation erscheint mit freundlicher Genehmigung von Royal Society of Chemistry. Reprinted with kind permission from Royal Society of Chemistry.

Erklärungen zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation

Optimization of Raman-spectrum baseline correction in biological application, S. Guo[1], T. Bocklitz[2], and J. Popp[3], *Analyst*, 2016, 141, 2396-2404.			
Beteiligt an (*Zutreffendes ankreuzen*)			
	1	2	3
Konzeption des Forschungsansatzes	x	x	x
Planung der Untersuchungen	x	x	x
Datenerhebung			
Datenanalyse und -interpretation	x	x	
Schreiben des Manuskripts	x	x	x
Vorschlag Anrechnung Publikationsäquivalente	1.0		

Volume 141 | Number 8 | 21 April 2016 | Pages 2321–2608

Analyst

www.rsc.org/analyst

ISSN 0003-2654

ROYAL SOCIETY OF CHEMISTRY

PAPER
Thomas Bocklitz, Jürgen Popp *et al.*
Optimization of Raman-spectrum baseline correction in biological application

175 YEARS

Analyst

PAPER

Cite this: *Analyst*, 2016, **141**, 2396

Optimization of Raman-spectrum baseline correction in biological application†

Shuxia Guo,[a] Thomas Bocklitz*[a,b] and Jürgen Popp[a,b,c]

In the last decade Raman-spectroscopy has become an invaluable tool for biomedical diagnostics. However, a manual rating of the subtle spectral differences between normal and abnormal disease states is not possible or practical. Thus it is necessary to combine Raman-spectroscopy with chemometrics in order to build statistical models predicting the disease states directly without manual intervention. Within chemometrical analysis a number of corrections have to be applied to receive robust models. Baseline correction is an important step of the pre-processing, which should remove spectral contributions of fluorescence effects and improve the performance and robustness of statistical models. However, it is demanding, time-consuming, and depends on expert knowledge to select an optimal baseline correction method and its parameters every time working with a new dataset. To circumvent this issue we proposed a genetic algorithm based method to automatically optimize the baseline correction. The investigation was carried out in three main steps. Firstly, a numerical quantitative marker was defined to evaluate the baseline estimation quality. Secondly, a genetic algorithm based methodology was established to search the optimal baseline estimation with the defined quantitative marker as evaluation function. Finally, classification models were utilized to benchmark the performance of the optimized baseline. For comparison, model based baseline optimization was carried out applying the same classifiers. It was proven that our method could provide a semi-optimal and stable baseline estimation without any chemical knowledge required or any additional spectral information used.

Received 6th January 2016,
Accepted 17th February 2016
DOI: 10.1039/c6an00041j
www.rsc.org/analyst

Introduction

Raman-spectroscopy is a non-invasive, label-free technique which reveals molecular fingerprints based on vibrational information. Raman-spectroscopy possesses properties, like its insensitivity to water, making it an ideal method for biological applications. Further advantages are the high spatial and temporal resolution, which can be easily realized.[1–9] Due to an improvement of the measurement equipment in the last decade Raman-spectroscopy has been widely applied for medical diagnosis,[4–6] for bacteria identification,[9] for tissue injury detection,[7] and even for surgical resection and decision making.[8]

However, Raman-spectroscopic measurements result in a mixture Raman-spectrum from all components, which feature a Raman-resonance within the laser focus. Because biological samples are heterogeneous mixtures of different biomolecules, biological Raman-spectra have a complex structure.[6,8] This complex structure reflects the biochemical composition of the sample. If Raman-spectra of different biological states are measured and compared, only subtle spectral differences are visible. A manual differentiating and rating of these subtle changes is not possible or practical. To overcome this issue, chemometrical methods and statistical models are applied if Raman-spectroscopy should be utilized for biological applications.[6,10]

In order to apply chemometrical methods a careful spectral pre-processing is necessary, which should remove corrupting effects and standardize the measured Raman-data. Within this pre-processing disturbing effects like comic spikes, white noise and baseline have to be removed. The baseline removal is an important correction, because the baseline is a few orders more intense as compared with Raman-bands.[11] If the baseline contribution is not properly corrected for, the resulting artefacts may hinder further data analysis. Such baseline artefacts would result in quite low generalization performances of the classification and regression models.

There are two basic approaches for baseline rejection, a physically and a mathematically motivated approach. Physical methods such as sample purification, shifted excitation, picosecond pulsing and gating require instrumental modifications.

---

[a] *Institute of Physical Chemistry and Abbe School of Photonics, Friedrich-Schiller-University, Jena, Helmholtzweg 4, D-07743 Jena, Germany*
[b] *InfectoGnostics Research Campus Jena, Centre of Applied Research, Philosophenweg 7, D-07743, Jena, Germany*
[c] *Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07702 Jena, Germany. E-mail: thomas.bocklitz@uni-jena.de*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6an00041j

This journal is © The Royal Society of Chemistry 2016

Mathematical methods, including derivative calculation, polynomial fitting, and frequency-domain filtering, do not need hardware modifications and therefore have been widely utilized. Recently, two automatic methods were proposed, which are based on iterative reweighted quantile regression[12] and iterative exponential smoothing,[13] respectively. However, none of these computational procedures can exactly separate the baseline from the measured Raman-spectrum. This results from the fact that the sum of the two quantities, the Raman-spectrum and the baseline, is measured during Raman-spectroscopic measurements. Usually, the baseline correction procedures estimate the baseline and subtract this estimation from the measured Raman-spectrum.[11] Thus the estimating procedure plays an important role for the performance of a baseline correction. In practice, analysts need to manually select an estimation method and choose its optimal parameters each time when they want to analyse a new dataset. This work is tedious, time-consuming, and experience dependent. Thus, an automatic optimization method is required, which performs the selection of estimation methods and parameters.

As it was reviewed in ref. 14, an automatic spectral pre-processing optimization can usually be realized by two approaches including the 'trial and error' method and a quality parameter based method. The former approach was applied in ref. 15 and 16, in order to optimize the pre-processing procedures, based on the output of regression and classification models.[15,16] In order to apply this scheme supervised statistical models are needed. It needs more data because an additional optimization of the statistical models is carried out. With the quality parameter based method no statistical models are required. Instead, a quality parameter is defined to evaluate the performance of the pre-processing procedures. By now parameters such as the 'simplicity value', the 'Pearson correlation coefficient', and the 'peak factor' have been employed in chromatographic and NMR (nuclear magnetic resonance) data processing. However, none of them performed well for the mid-infrared data.[14] The so-named 'super parameter' was utilized for NIR (near-infrared) spectra by combining the explained variance of PC1 (the first principal component), number of outliers, and a coefficient of variation.[17] Within this scheme, several numeric thresholds are required, which have to be adapted and therefore specific prior spectral information is needed.

To the authors' knowledge, no similar parameters are used for Raman-spectral pre-processing. Therefore a method is needed, which allows to optimize the Raman-baseline correction only using spectral features within the Raman-spectrum itself. Therefore, we proposed an approach to automatically optimize baseline correction given a certain dataset. We decided to investigate the baseline correction separately, as it has the highest impact on the chemometrical analysis.

This manuscript is structured into three sections. First, a quantitative marker for the quality of baseline estimations was defined. Through a grid search process, every combination of a baseline estimation method with its parameters was tested and the defined marker was calculated. Thereafter we built an optimization framework based on a genetic algorithm (GA), with the defined marker as evaluation function, and the combination of a method with its parameters as the chromosome. Finally, classifiers based on a principal component analysis (PCA) and support vector machine (SVM) applying the linear and radial kernel were utilized. Raman-spectra, which were corrected with our optimized baseline estimation, were analysed with these classifiers. The mean sensitivities were investigated to test the performance of our optimized baseline estimation. For comparison the baselines featuring the highest mean sensitivities were inspected and compared with the results of our proposed method.

Experimental

Raman-spectroscopy

The Raman-spectroscopic measurements were published by our group in ref. 10, which would be only briefly summarized in this section. Breast carcinoma derived tumor cells (MCF-7, BT-20) and acute myeloid leukemia cells (OCI-AML3) were grown. Raman-spectra were measured with an excitation wavelength of 785 nm (model xtra, Toptica, Germany) under an upright Raman-microscope (Microprobe, Kaiser Optical Systems, USA) with a 60×/NA 1.0 water immersion objective (Nikon, Japan) and 75 mW power at the sample. In total, 1553 cells (558 MCF-7, 477 BT-20, 518 OCI-AML3) were measured.

Computation

All computations were done in statistical programming language Gnu R.[18] The packages 'signal',[19] 'Peaks',[20] 'baseline',[21] 'simecol',[22] 'genalg'[23] and 'e1071'[24] were utilized. The functions from the packages were complemented by in-house written procedures.

Data analysis

Baseline correction. In this contribution, three baseline correction methods were applied, including sensitive nonlinear iterative peak (SNIP) clipping,[20] asymmetric least squares (ALS), and modified polynomial (Modpoly) fitting.[21] The SNIP gradually clips out a Raman-peak region by replacing its values with the minima within this region. The ALS estimates baselines by a combination of an iterative least squares smoothing with asymmetric weights for positive and negative intensity values. The Modpoly produces a baseline by a polynomial fitting based on the original Raman-spectrum and an iterative procedure. All these methods depend on the values of their

Table 1 Baseline methods and parameters

Method	Para1	Range	Para2	Range
SNIP	order	'2', '4'	iterations	[1,100]
ALS	λ	[3.5,10.5]	p	[0.001,0.1]
Poly	degree	[3,10]	/	/

parameters. Thus their parameters are tabled and the ranges, which were investigated, are summarized in Table 1.

$$\begin{aligned} t &= \frac{\max(I_z - \min(I_z))}{\mathrm{mean}(I_n)} \\ m_1 &= \frac{\log(n_p)}{A_p} + \frac{A_z t}{\log(n_z)} \\ m_2 &= \frac{A_p}{(A_z t + A_p)} \\ R^{12} &= m_1/m_2 \end{aligned} \tag{1}$$

Quantitative marker. To define a reasonable marker for the baseline estimation's quality a few considerations about features of an optimal baseline are necessary. A baseline correction should result in a small intensity loss for regions featuring Raman-peaks. Also should the corrected Raman-spectrum be flat without artefact peaks for regions where no Raman-information is expected. Thus an optimal baseline correction should provide a maximal A_p/A_z, the ratio of peak area to area of the region with no Raman-information. However, it is not straightforward to automatically optimize baseline correction according to A_p/A_z. During an iterative peak clipping process, the value of A_p/A_z is quite small in the beginning and it increases when the estimated baseline is becoming flatter. However, the value of A_p/A_z may be almost stable from some point on, even if the baseline estimation becomes flatter as required. This makes the automatic optimization rather challenging since a too flat baseline estimation would be selected by maximizing A_p/A_z. To deal with this issue, we defined a quantitative marker for a baseline estimation as formula (1), where the terms I, A, n denote the Raman-intensity, the area, and the number of wavenumber positions, respectively. The subscripts p, z, and n represent Raman-peaks positions, positions without Raman-bands and regions for spectral normalization.

Within these definitions, the term t represents the normalized difference between the maximum and minimum within the no-Raman-information region. Here a region is selected and the mean intensity of this region ($\mathrm{mean}(I_n)$) is used as a normalization term. This region can be the combination of the regions with and without expected Raman-peaks. t is calculated as a penalty for negative values in a Raman-spectrum after baseline correction. A larger t is expected if such negative values exist. The term m_1 is defined as a sum of the inverse mean intensity of the peak region and the mean intensity of the no-information region. We used the mean intensity instead of the area in case the area of peak regions dominates the area of no-Raman-information region. The mean intensity of the peaks region is inversed for the first term to make sure m_1 will become a minimum for an optimal baseline correction. Here A_z is multiplied by t, thus the second term becomes larger, if the aforementioned negative values exist. To compensate the large difference between the amplitudes of n_p and n_z in comparison with A_p and A_z, the logarithm is computed. Otherwise m_1 would be dominated by n_p if it is large. Generally, for a Raman-spectrum, after a perfect baseline correction, the area where no Raman-information is expected (A_z) should be minimal, while the area where Raman-peaks might occur (A_p) should be maximal. Therefore, m_1 is expected to be the minimum for a good baseline estimation. The term m_2 represents the proportion of A_p to the sum ($A_z t + A_p$). A_z is also multiplied by t as a penalty for negative values. m_2 is expected larger for better baseline corrections and should be close to one in the best cases. R^{12}, the ratio of m_1 to m_2, is the final quantitative marker and should be minimal for an optimal baseline estimation.

To validate this definition and our hypothesis, we performed a simulation on three artificial spectra, constructed with eight exactly known peaks and three different baseline profiles. The baseline correction was optimized according to two mechanisms, (1) the minimal R^{12} value, (2) the minimal Euclidean distance between the reference (true) spectrum and the corrected spectrum. As is shown in ESI,† we obtained quite consistent baseline correction results by these two mechanisms. That is to say, the baseline correction giving the minimal R^{12} value is reasonable, with ignorable over- or under-corrections.

Before the value of R^{12} can be calculated, spectral processing should be carried out, as is shown in Fig. 1. The processing steps include calibration (block in yellow), pre-processing (blocks in green), and finding local minima within the Raman-spectra (block in blue). The last step is necessary for the definition of peak and no-peak regions. It can be done automatically, like we describe it here, or by manual definition in order to incorporate spectral expert knowledge.

First of all, the wavenumber of all Raman-spectra were calibrated using the CaF_2 peak at 322 cm^{-1} and all Raman-spectra were interpolated to an equidistant wavenumber grid of 1 cm^{-1}.[10,25] The mean Raman-spectra of each cell type (MCF-7, BT-20, OCI-AML3) were calculated and plotted in Fig. 2. All computations afterwards were carried out with the help of these three mean Raman-spectra.

For pre-processing, white noise was removed by a Savitzky-Golay filtering with a window width of 11 and an order of 2. Then the baseline correction was carried out with a combination of a baseline estimation method and its parameters

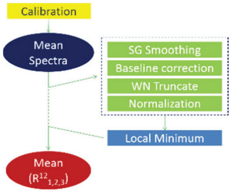


Fig. 1 Workflow of the R^{12} calculation: The mean Raman-spectrum of each cell type is computed after wavenumber calibration. Afterwards, the mean Raman-spectrum is pre-processed including smoothing, baseline correction, wavenumber truncation and normalization. Finally, local minima within the Raman-spectra were searched and utilized as regions, which exhibit no Raman-information.

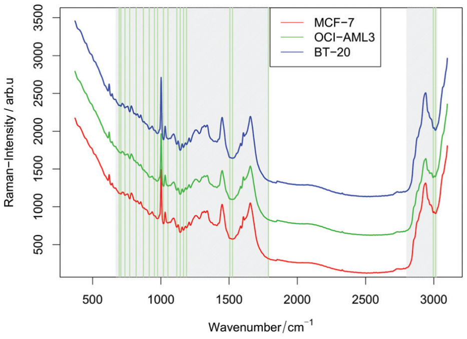


Fig. 2 Mean Raman-spectra of the three investigated cell types, which were used for R^{12} calculation. The grey and green shades mark the Raman-peak area and regions not containing Raman-information, respectively. Here both regions were determined automatically by local minimum searching.

listed in Table 1. Afterwards, the baseline corrected mean Raman-spectra were truncated to the wavenumber regions from 675 to 1785 cm^{-1} and from 2815 to 3020 cm^{-1}.[10] Finally, the Raman-spectra were normalized based on the spectral region from 675 to 1785 cm^{-1}, which was also selected as the normalization region for term t in formula (1). Within this region, even bad baseline estimations give acceptable correction results. This is important as the spectral normalization is not introducing a strong bias. This is not the case if the region from 2815 to 3020 cm^{-1} is also used for the normalization.

After the pre-processing was carried out, wavenumber positions featuring local minima of the Raman-intensity have to be found. This can be done automatically or by manual selection. In this contribution, an automatic searching algorithm was applied, from which the results were shown in green vertical lines in Fig. 2.

The found wavenumber positions were considered as positions where no Raman-information is present. The wavenumber positions not belonging to the defined no Raman-information region were defined as Raman-peak region (shown in grey shades). According to formula (1), the marker R^{12} was calculated for all three mean Raman-spectra, resulting in three values. The averaged value was stored as the final result of the quantitative marker for a certain combination of baseline estimation methods and parameters.

In order to check the influence of the estimation algorithms and the parameters on the marker R^{12}, a grid search was utilized. It calculates the marker R^{12} for every combination of the three mentioned baseline correction methods with their parameters. The corresponding searching ranges are listed in Table 1. R^{12} was calculated for each combination. Therefore, the marker R^{12} was known for all combinations.

Baseline optimization. Within a grid search we have tried all possible combinations of baseline correction methods and the corresponding parameters. Theoretically the optimal baseline correction can be obtained by grid search. For practical applications, however, this is not reasonable, as it is time-consuming and needs a lot of calculation power. Therefore we used a genetic algorithm to optimize the baseline correction without trying all possibilities.

Genetic algorithm, a special evolutionary algorithm, starts from a random population of solutions and optimizes the solutions according to Charles Darwin's evolution theory. Firstly, an evaluation function needs to be defined as a criterion of fitness. Secondly, a number of combinations of parameters, namely chromosomes, have to be created. Afterwards, values of the defined evaluation function from each chromosome are calculated. According to these values the chromosomes are evolved by selection, crossover, mutation, and dying in order to generate a new generation. This new generation is tested applying the evaluation function and evolved again. Genetic algorithm keeps this iteration of evolution-and-test to search better chromosomes with a better value of evaluation function. The algorithm stops if some termination conditions are reached (Fig. 3). The best chromosome in the last generation is selected as the optimum.[26]

In our investigation, the target of the optimization was a combination of a baseline estimation method and its parameters. The optimal baseline estimation can be determined by the minimal R^{12} value. Thus the genetic algorithm was established with the quantitative marker R^{12} as the evaluation function. Chromosomes were composed of three genes. The first gene represents the index of baseline correction methods. The second gene represents the parameter 'order' for SNIP, 'λ' for ALS, and 'degree' for Modpoly. The third gene represents 'iterations' for SNIP and 1000 times of 'p' for ALS. It is ignored in the case of the Modpoly method. Chromosomes and the evolution ranges for each gene are summarized in Table 1 and Fig. 3.

The function 'rbga' in the package 'genalg' was used with the population size of 5, the mutation rate of 1/6 and the

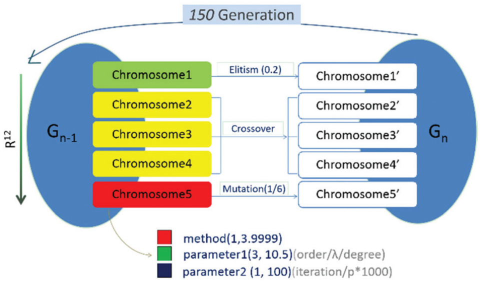


Fig. 3 Flowchart of the genetic algorithm: Three genes within the chromosomes were applied, representing the baseline correction method and its two parameters, respectively. For the Modpoly method the third gene was ignored. A generation with population size of 5 was randomly created in the beginning. The algorithm developed for 150 generations based on the evolutional theory. If a chromosome had lower R^{12} values, the probability to survive increases.

elitism rate of 20%. The algorithm evolved for 150 generations before it is terminated.

Classification. To check the performance of the optimal baseline correction, we applied statistical models to classify the Raman-spectra of three cell types (MCF-7, BT-20 and OCI-AML3). Firstly the dimension of the dataset was reduced by a principal component analysis with the first 40 PCs kept. The number was chosen quite high in order to avoid an indirect correction of baseline drifts. Afterwards classifiers based on a support vector machine were constructed, applying a linear and a radial kernel function, respectively. A batch-out cross-validation was used, with Raman-spectra from each batch taken out once and the averaged mean sensitivities were stored.

Results and discussion

Grid search

The result of grid search is visualized in false-colours in Fig. 4a and the five lowest R^{12} values are marked. Their corresponding baseline estimations are plotted in Fig. 4b together with the raw spectrum. It can be seen that these five baselines are looking reasonable and have the least intensity losses within the Raman-peak region. Besides this, the baseline with the maximal R^{12} value and two baselines with moderate R^{12} values are plotted in Fig. 4c. The baseline corresponding to the maximal R^{12} value almost clipped out all the Raman-peaks, while the two baselines with the moderate R^{12} values were either under-fitted or over-fitted. Particularly, for an over-fitted baseline estimation, for instance, the red and green baselines in Fig. 4c, the first term of m_1 in eqn (1) would be much larger than the second term. On the contrary, for an under-fitted baseline estimation such as the blue baseline in Fig. 4c, the first term of m_1 in eqn (1) would be quite smaller than the second term. In both over- and under-fitted cases, m_1 is not the minimum and a large R^{12} value is generated. Besides, m_2 decreases for under-fitted baseline estimations, producing a higher R^{12}. The optimal baseline can be found only if both terms of m_1 in eqn (1) are small while m_2 is close to one. This behaviour indicates the good performance of the defined quantitative marker. Specifically, the five baseline estimations corresponding to the five minimal R^{12} values are quite similar thus any of them can be used. Hence it is not necessary to search for the exactly lowest R^{12}, which is actually not possible in practice.

Genetic algorithms

After carrying out the genetic algorithm, the best chromosome in every generation and its value of R^{12} are plotted in Fig. 5a and b. The value of R^{12} decreased from 5.93 in the first generation to 5.24 in the last generation. From Fig. 5a, it can be concluded that the value of R^{12} converged within fifty generations. This behaviour is obvious in Fig. 5c. The optimal baselines within the first and last generation are plotted in Fig. 5d. The baseline in the first generation was under-fitted in the spectral regions of 675–1214 cm^{-1} and 2815–3020 cm^{-1}. This results in an artificial envelope and an increase of the Raman-peaks after baseline correction. Nonetheless, the baseline in the last generation fits better within the whole spectral region of interest (675–1785 cm^{-1}, 2815–3020 cm^{-1}). It is noteworthy that after optimization, satisfactory baseline corrections were observed for peaks with various widths. This means with our method baseline correction parameters can be well balanced for wide and narrow peaks. Moreover, the optimal baseline in the fiftieth generation (Fig. 5d) is quite similar to the one of the last generation, proving the convergence of R^{12} value within fifty generations. As shown in Fig. 5e, the best baseline optimized by the genetic algorithm was comparable to the one from grid search, indicating a good performance of the genetic algorithm based method. However, with this method, the optimization was faster without trying all combinations of baseline correction method with its parameters. An advantage of the GA method is that an optimization of not only three but also much more baseline correction methods and their respective parameters is possible. Furthermore an

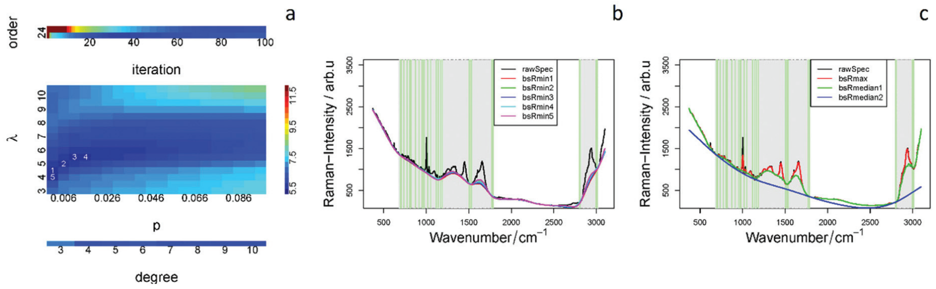


Fig. 4 Results of a grid search: (a) R^{12} results from grid search of the three-background estimation methods and their parameters. All values higher than twelve were clipped to twelve allowing a convenient visualization. The first five minimal R^{12} values are marked. (b) Baseline estimations corresponding to the lowest five R^{12} values are plotted. (c) Baseline estimates with the maximum R^{12} value and two moderate R^{12} values are shown. These estimations are either over-fitted or under-fitted.

This journal is © The Royal Society of Chemistry 2016

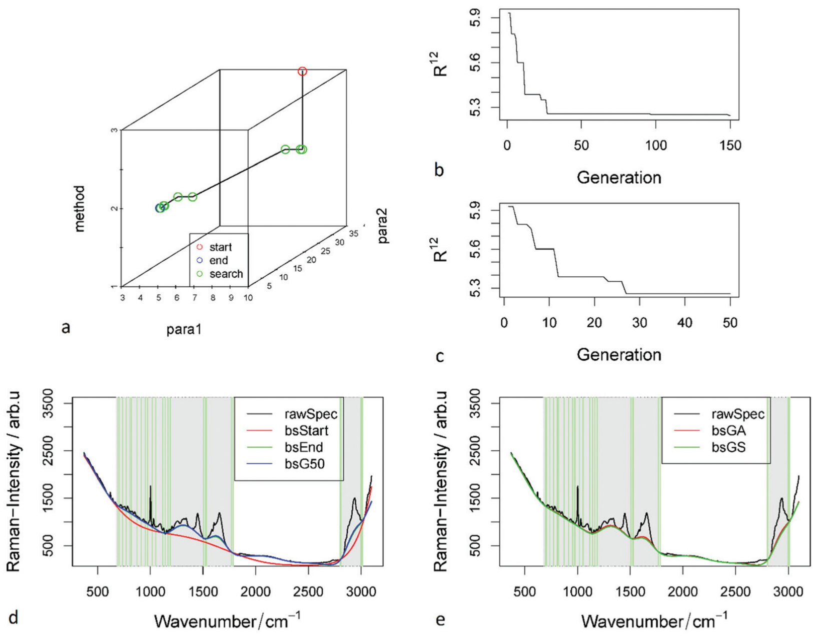


Fig. 5 Results from the genetic algorithm: (a) plot of the values of the genes from the best chromosome through 150 generations are visualized. The algorithm started from the points in red and developed until termination at the point in blue. (b) R^{12} value from the first to 150th generation, which decreased from 5.93 to 5.24, is plotted. (c) R^{12} value within the first to 50th generation is visualized, indicating the algorithm converged within 50 generations. (d) Baselines from 1st, 50th and 150th generation are depicted. The baseline estimation was under-fitted in the beginning (1st generation). The baseline from the 50th generation was similar to the one in the last generation (150th), indicating the convergence within 50 generations. (e) The optimal baseline estimations through grid search and the genetic algorithms are comparable. Their similarity prove the good performance of the genetic algorithm based method.

effective searching of good baseline methods for large-size datasets can be achieved.

Validation and discussion

So far, a marker for quantifying the quality of a baseline estimation has been defined and is proven to perform well. Now we investigate if the optimal baseline estimation with respect to the minimal R^{12} features optimal classification results as well. In order to test this, a classification was carried out after the whole pre-processing was performed and the mean sensitivity was calculated to check the performance. A grid search was carried out again and the baseline correction with the best classification result is selected as the benchmark. The corresponding mean sensitivity was compared with the value resulting from the baseline correction optimized with the GA based methodology.

The results are plotted in Fig. 6a, where the R^{12} values of the grid search were converted to a vector and plotted in a decreasing order. Correspondingly, the mean sensitivities of the both classifiers were plotted. It is obvious that the mean sensitivities of both classifiers were increasing when R^{12} decreased. The highest mean sensitivities were 74.3% and 76.8% for the linear and radial kernel SVM, respectively. While the respective mean sensitivities were 72.7% and 73.1%, if our optimized baseline estimation is applied. Despite of the small differences, the classification results from our optimized baseline is comparable with the highest mean sensitivities. Remarkably, our method prevents operators from selecting a bad baseline estimation and not optimal parameters.

The baseline estimations with the highest mean sensitivities were shown in Fig. 6b. In comparison the optimal baseline estimation based on our method is plotted as well. Apparently, the two baselines with the highest mean sensitivity feature a higher intensity loss within the Raman-peak regions compared with our optimized baseline. That is to say, despite of the highest mean sensitivity, the baselines are from a spectroscopic point view sub-optimal. Furthermore, the mean sensitivities from the two classifiers are similar but not identical, leading to slightly different selected baselines. Even though not strongly proved in this experiment, it could be deduced that different baselines are selected, if different statistical models or classifiers are used. This means the selected baseline is model dependent, which is a drawback.

This journal is © The Royal Society of Chemistry 2016

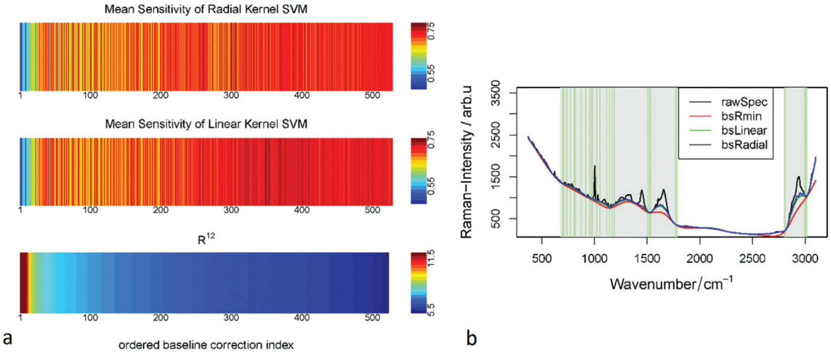


Fig. 6 Results of three-class classification: (a) R^{12} and mean sensitivities from both classifiers are visualized. The distribution of mean sensitivities from both classifiers differed slightly. (b) Baselines with the highest mean sensitivity for both classifiers are visualized. Compared to the baseline estimation with the lowest R^{12} value (in red), they featured a higher intensity loss in the peak region.

Besides the above mentioned three-class classification task, a two-class (MCF-7, OCI-AML3) classification task was carried out using PCA (with 40 PCs kept) and linear kernel SVM. The results are shown in Fig. 7. Here the baseline with the highest mean sensitivity (88.6%) also resulted in the lowest R^{12} value. Nonetheless the baselines with the second and third highest mean sensitivity featured a high R^{12}. This means even bad baselines would possibly produce very good classification results. It is probably because not well adapted baseline corrections would introduce additional information such as artificial envelopes and increase the differences between Raman-spectra. This may also be the case if residual fluorescence is used for separating the classes. It may be concluded from this fact that model based optimization of the baseline correction has to be used with caution.

Moreover, comparing the baseline results shown in Fig. 6b and 7b, the same baseline estimation was obtained for the two classification tasks with our proposed method. It is not the case for the model based optimization approach. Therefore, the latter optimization method is not only model but also task dependent.

To summarize, the baseline estimation selected according to the highest classification mean sensitivity is model and task dependent and is not always a good choice. Besides this, it is computationally expensive and time-consuming because for every combination of baseline estimation method and parameters a statistical model has to be constructed. In contrast our proposed method could produce a stable and reliable baseline with comparable classification results. As shown in Fig. S2† the baseline optimized by our method is close to the true baseline for a simulated case. These simulations strongly indicate that an application is also possible in quantitative analyses. Furthermore, the proposed method prevents scientists from selecting a not optimal baseline as an over-correction and under-correction can be avoided. Last but not the least its computational expense is much less since no additional statistical models have to be constructed.

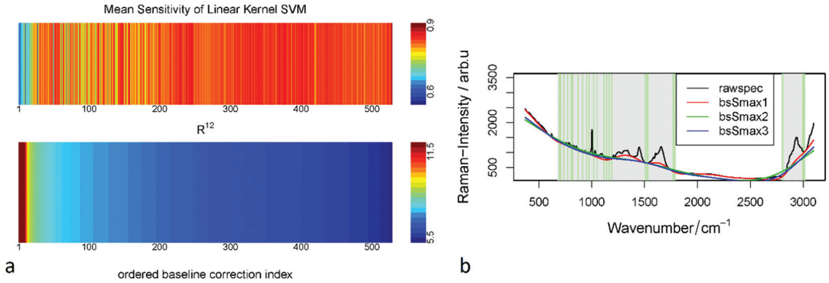
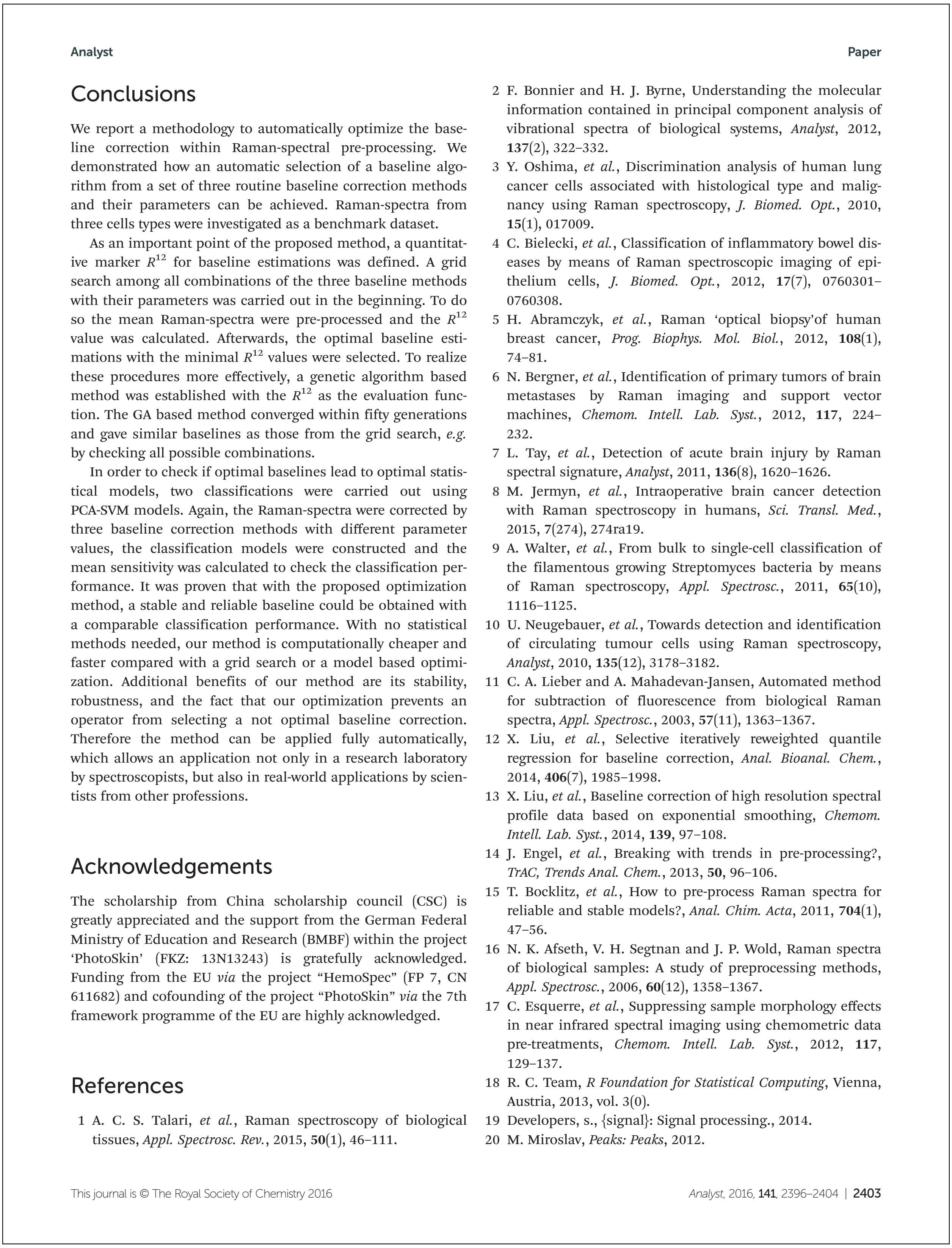


Fig. 7 Results of two-class classification: (a) R^{12} and mean sensitivity for a linear SVM classifier is plotted. (b) Baseline with the highest mean sensitivity also gave the lowest R^{12}. However, the baselines with the second and third highest mean sensitivities featured high R^{12} values. This demonstrates the instability of baselines selected according to the highest mean sensitivity. The baseline estimations with the lowest R^{12} in Fig. 6b and 7b are the same, which indicates the stability of our method and our optimization is not task dependent.

This journal is © The Royal Society of Chemistry 2016



Conclusions

We report a methodology to automatically optimize the baseline correction within Raman-spectral pre-processing. We demonstrated how an automatic selection of a baseline algorithm from a set of three routine baseline correction methods and their parameters can be achieved. Raman-spectra from three cells types were investigated as a benchmark dataset.

As an important point of the proposed method, a quantitative marker R^{12} for baseline estimations was defined. A grid search among all combinations of the three baseline methods with their parameters was carried out in the beginning. To do so the mean Raman-spectra were pre-processed and the R^{12} value was calculated. Afterwards, the optimal baseline estimations with the minimal R^{12} values were selected. To realize these procedures more effectively, a genetic algorithm based method was established with the R^{12} as the evaluation function. The GA based method converged within fifty generations and gave similar baselines as those from the grid search, *e.g.* by checking all possible combinations.

In order to check if optimal baselines lead to optimal statistical models, two classifications were carried out using PCA-SVM models. Again, the Raman-spectra were corrected by three baseline correction methods with different parameter values, the classification models were constructed and the mean sensitivity was calculated to check the classification performance. It was proven that with the proposed optimization method, a stable and reliable baseline could be obtained with a comparable classification performance. With no statistical methods needed, our method is computationally cheaper and faster compared with a grid search or a model based optimization. Additional benefits of our method are its stability, robustness, and the fact that our optimization prevents an operator from selecting a not optimal baseline correction. Therefore the method can be applied fully automatically, which allows an application not only in a research laboratory by spectroscopists, but also in real-world applications by scientists from other professions.

Acknowledgements

The scholarship from China scholarship council (CSC) is greatly appreciated and the support from the German Federal Ministry of Education and Research (BMBF) within the project ‘PhotoSkin’ (FKZ: 13N13243) is gratefully acknowledged. Funding from the EU *via* the project “HemoSpec” (FP 7, CN 611682) and cofounding of the project “PhotoSkin” *via* the 7th framework programme of the EU are highly acknowledged.

References

1 A. C. S. Talari, *et al.*, Raman spectroscopy of biological tissues, *Appl. Spectrosc. Rev.*, 2015, **50**(1), 46–111.

2 F. Bonnier and H. J. Byrne, Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems, *Analyst*, 2012, **137**(2), 322–332.

3 Y. Oshima, *et al.*, Discrimination analysis of human lung cancer cells associated with histological type and malignancy using Raman spectroscopy, *J. Biomed. Opt.*, 2010, **15**(1), 017009.

4 C. Bielecki, *et al.*, Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells, *J. Biomed. Opt.*, 2012, **17**(7), 0760301–0760308.

5 H. Abramczyk, *et al.*, Raman ‘optical biopsy’of human breast cancer, *Prog. Biophys. Mol. Biol.*, 2012, **108**(1), 74–81.

6 N. Bergner, *et al.*, Identification of primary tumors of brain metastases by Raman imaging and support vector machines, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 224–232.

7 L. Tay, *et al.*, Detection of acute brain injury by Raman spectral signature, *Analyst*, 2011, **136**(8), 1620–1626.

8 M. Jermyn, *et al.*, Intraoperative brain cancer detection with Raman spectroscopy in humans, *Sci. Transl. Med.*, 2015, 7(274), 274ra19.

9 A. Walter, *et al.*, From bulk to single-cell classification of the filamentous growing Streptomyces bacteria by means of Raman spectroscopy, *Appl. Spectrosc.*, 2011, **65**(10), 1116–1125.

10 U. Neugebauer, *et al.*, Towards detection and identification of circulating tumour cells using Raman spectroscopy, *Analyst*, 2010, **135**(12), 3178–3182.

11 C. A. Lieber and A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, *Appl. Spectrosc.*, 2003, **57**(11), 1363–1367.

12 X. Liu, *et al.*, Selective iteratively reweighted quantile regression for baseline correction, *Anal. Bioanal. Chem.*, 2014, **406**(7), 1985–1998.

13 X. Liu, *et al.*, Baseline correction of high resolution spectral profile data based on exponential smoothing, *Chemom. Intell. Lab. Syst.*, 2014, **139**, 97–108.

14 J. Engel, *et al.*, Breaking with trends in pre-processing?, *TrAC, Trends Anal. Chem.*, 2013, **50**, 96–106.

15 T. Bocklitz, *et al.*, How to pre-process Raman spectra for reliable and stable models?, *Anal. Chim. Acta*, 2011, **704**(1), 47–56.

16 N. K. Afseth, V. H. Segtnan and J. P. Wold, Raman spectra of biological samples: A study of preprocessing methods, *Appl. Spectrosc.*, 2006, **60**(12), 1358–1367.

17 C. Esquerre, *et al.*, Suppressing sample morphology effects in near infrared spectral imaging using chemometric data pre-treatments, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 129–137.

18 R. C. Team, *R Foundation for Statistical Computing*, Vienna, Austria, 2013, vol. 3(0).

19 Developers, s., {signal}: Signal processing., 2014.

20 M. Miroslav, *Peaks: Peaks*, 2012.

21 K. H. Liland and B. H. Mevik, *Baseline: Baseline Correction of Spectra.*, 2015.

22 T. Petzoldt and K. Rinke, Simecol: an object-oriented framework for ecological modeling in R, *Journal of Statistical Software*, 2007, **22**(9), 1–31.

23 E. Willighagen and M. Ballings, *Genalg: R Based Genetic Algorithm.*, 2015.

24 E. Dimitriadou, *et al.*, Misc functions of the Department of Statistics (e1071), TU Wien., *R package version*, 2014, 1.6-4.

25 R. L. McCreery, *Raman spectroscopy for chemical analysis*, John Wiley & Sons, 2000, vol. 157.

26 C. B. Lucasius and G. Kateman, Understanding and using genetic algorithms Part 1. Concepts, properties and context, *Chemom. Intell. Lab. Syst.*, 1993, **19**(1), 1–33.

This journal is © The Royal Society of Chemistry 2016

Journal Name

ARTICLE

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Optimization of Raman-Spectrum Baseline Correction in Biological Application

Shuxia Guo,[a] Thomas Bocklitz,[a,b,*] and Jürgen Popp[a,b,c]

In the last decade Raman-spectroscopy has become an invaluable tool for bio-medical diagnostics. However, a manual rating of the subtle spectral differences between normal and abnormal disease states is not possible or practical. Thus it is necessary to combine Raman-spectroscopy with chemometrics in order to build statistical models predicting the disease states directly without manual intervention. Within chemometrical analysis a number of corrections have to be applied to receive robust models. Baseline correction is an important step of the pre-processing, which should remove spectral contributions of fluorescence effects and improve the performance and robustness of statistical models. However, it is demanding, time-consuming, and depends on expert knowledge to select an optimal baseline correction method and its parameters every time working with a new dataset. To circumvent this issue we proposed a genetic algorithm based method to automatically optimize the baseline correction. The investigation was carried out in three main steps. Firstly, a numerical quantitative marker was defined to evaluate the baseline estimation quality. Secondly, a genetic algorithm based methodology was established to search the optimal baseline estimation with the defined quantitative marker as evaluation function. Finally, classification models were utilized to benchmark the performance of the optimized baseline. For comparison, model based baseline optimization was carried out applying the same classifiers. It was proven that our method could provide a semi-optimal and stable baseline estimation without any chemical knowledge required or any additional spectral information used.

Simulation

Spectra construction

To allow an understanding of the working of our proposed method artificial spectra were constructed within the wavenumber range from 300 to 3000 cm^{-1}.

Firstly, we combined eight Gaussian peaks at arbitrary wavenumber positions 500, 700, 900, 970, 1050, 1500, 2000 and 2600 cm^{-1}. The maximum intensities of these peaks varied within the range 800-2200, while the full width at half maximum (FWHM) varied within the interval 50-120 cm^{-1}. The three peaks at 900, 970, and 1050 cm^{-1} were chosen so close that they overlapped to generate a complex structure (See Fig. S1 (a)).

Secondly, three series of curves were prepared to construct baseline profiles. Each series contains a second-order polynomial and five Gaussian peaks with large bandwidth. Details about the related parameters are listed in Tab. S1. Accordingly, three baseline profiles were created by adding up all compositions within each curve series, which are plotted in Fig. S1 (b).

Finally, the pure spectrum (Fig. S1(a)) was added up with the three baseline profiles (Fig. S1(b)), generating three spectra. Additionally, Poisson distributed noise was generated and added up to these three spectra. In this way, three simulated spectra with different baseline patterns were created, as shown in Fig. S1 (c).

Tab. S1. Parameters of the three curve series for constructing baseline profiles

		Gaussian 1	Gaussian 2	Gaussian 3	Gaussian 4	Gaussian 5	2nd order Polynomial
Series 1	A	1500	-750	1125	1125	1875	$0.00015*(x-600)^2$
	μ	500	1000	1500	2300	2800	
	σ	900	900	900	900	900	
Series 2	A	2250	-375	-1125	1875	1500	$0.00015*(x-1500)^2$
	μ	400	800	1200	1800	2800	
	σ	1000	1000	1000	1000	1000	
Series 3	A	1875	-1125	1500	-1125	1875	$0.00015*(x-2500)^2$
	μ	600	1200	1700	2000	2700	
	σ	1100	1100	1100	1100	1100	

Grid search

All simulated spectra were smoothed by a Savitzky-Golay filtering with a window width of 11 and an order of 2. Reference spectra were obtained by subtracting the true baseline profiles from the smoothed spectra. The subtracted spectra were used as reference spectra instead of the pure spectrum shown in Fig. S1(a). This was done to make the reference spectra comparable to baseline corrected spectra, which were also generated from the smoothed spectra.

The same grid search procedure as for the real Raman spectra was performed on the smoothed spectra. The Euclidean distance

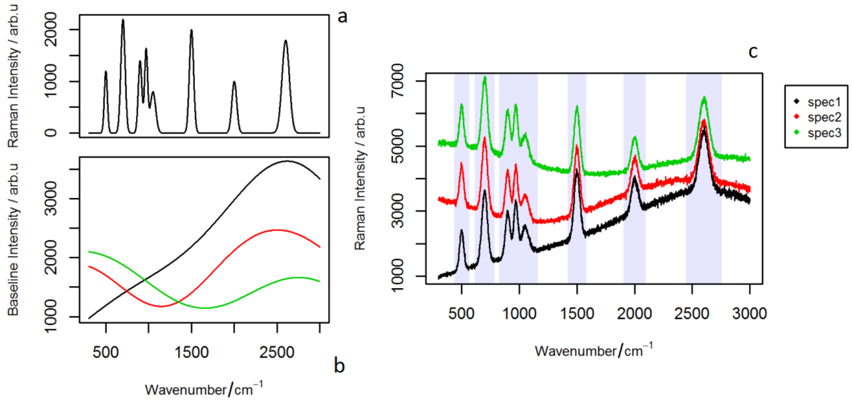


Fig. S1 Constructed Spectra for simulation. (a) Pure spectrum containing eight peaks with various intensities and FWHMs. A wide peak was constructed by an overlap of the three peaks at 900, 970, and 1050 cm^{-1}. (b) Three baseline profiles were generated by combination of one second order polynomial and five Gaussian peaks with large bandwidth. (c) Simulated spectra, created from three parts, pure spectra, baseline, and Poisson distributed noise.

between the baseline corrected spectra and the reference spectra ($\|S_c - S_r\|_2$) was computed and employed as the benchmark of baseline correction. Here we assume that a 'correct' baseline correction provide a minimal Euclidean distance. Meanwhile, the R^{12} values were calculated according to eq. (1). The grey shaded regions shown in Fig. S1(c) were used as peak region, while the rest was utilized as no-Raman-information region. All three Raman spectra were vector normalized within the whole wavenumber range. According to our purposed method, an optimal baseline correction can be expected at the minimum of R^{12}. Afterwards, the optimal baseline corrections were selected according to two mechanisms, i.e., the minimal R^{12} value and the minimal Euclidean distance. The results were plotted in Fig. S2, where the region with the overlapped peak was highlighted in a zoomed image. As shown for the first two simulated spectra, the two mechanisms yielded an identical baseline correction. While a slight difference was observed for the third simulated spectrum. Besides this neither of mechanism can exactly eliminate the baseline, demonstrated by the mismatch between the reference and the baseline corrected spectra. This also indicates the impossibility to exactly separate baselines and Raman signals. Nevertheless, a high consistency was observed between the baseline corrections optimized by these two mechanisms. That is to say, the R^{12} value can indeed reflect the goodness of baseline correction without knowing the reference spectra, which means a feasibility to optimize baseline correction according to R^{12} values.

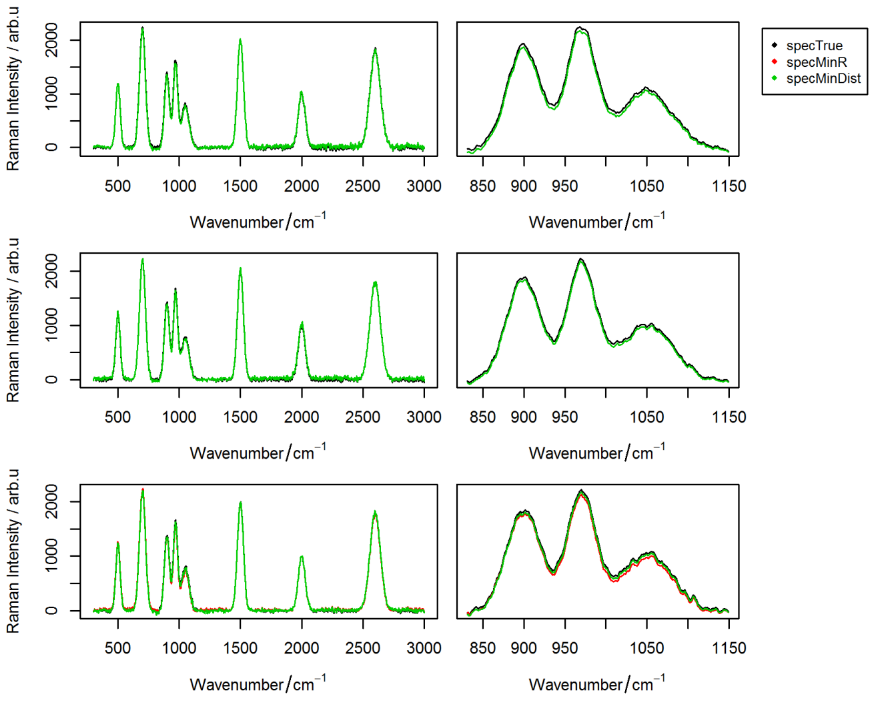


Fig. S2. The reference spectra (black) and optimal baseline corrections obtained by the minimal R^{12} value (red) and minimal Euclidean distance (green). The three overlapped peaks are highlighted in a zoomed version on the right. For the first two simulated spectra, the two mechanisms yielded the identical optimal baseline correction. A slight difference was observed for the third simulated spectrum. Nevertheless, a high consistency was observed between the baseline corrections optimized by the two mechanisms. Besides, a mismatch between the reference and the baseline corrected spectra is observed, indicating the impossibility to exactly separate baselines and Raman signals.

7.2 Spectral Reconstruction for Shifted-Excitation Raman Difference Spectroscopy (SERDS) (A2)

S. Guo, O. Chernavskaia, J. Popp, and T. Bocklitz, *Talanta*, 2018, 186, 372-380.

Der Nachdruck der folgenden Publikation erscheint mit freundlicher Genehmigung von ELSEVIER. Reprinted with kind permission from ELSEVIER.

Erklärungen zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation

Spectral reconstruction for shifted-excitation Raman difference spectroscopy (SERDS), S. Guo[1], O. Chernavskaia[2], J. Popp[3], and T. Bocklitz[4], *Talanta*, 2018, 186, 372-380.				
Beteiligt an (*Zutreffendes ankreuzen*)				
	1	2	3	4
Konzeption des Forschungsansatzes	x		x	x
Planung der Untersuchungen	x	x	x	x
Datenerhebung		x		
Datenanalyse und -interpretation	x	x		x
Schreiben des Manuskripts	x		x	x
Vorschlag Anrechnung Publikationsäquivalente	1.0			

Talanta 186 (2018) 372–380

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Spectral reconstruction for shifted-excitation Raman difference spectroscopy (SERDS)

Shuxia Guo[a,b], Olga Chernavskaia[a,b], Jürgen Popp[a,b,c], Thomas Bocklitz[a,b,*]

[a] Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany
[b] Institute of Physical Chemistry and Abbe Centre of Photonics, Friedrich-Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany
[c] InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, 07743 Jena, Germany

ARTICLE INFO

Keywords:
Raman spectroscopy
Shifted excitation Raman difference spectroscopy
Spectral reconstruction
Chemometrics
Fluorescence background

ABSTRACT

Fluorescence emission is one of the major obstacles to apply Raman spectroscopy in biological investigations. It is usually several orders more intense than Raman scattering and hampers further analysis. In cases where the fluorescence emission is too intense to be efficiently removed via routine mathematical baseline correction algorithms, an alternative approach is needed. One alternative approach is shifted-excitation Raman difference spectroscopy (SERDS), where two Raman spectra are recorded with two slightly different excitation wavelengths. Ideally, the fluorescence emission at the two excitations does not change while the Raman spectrum shifts according to the excitation wavelength. Hence the fluorescence is removed in the difference of the two recorded Raman spectra. For better interpretability a spectral reconstruction procedure is necessary to recover the fluorescence-free Raman spectrum. This is challenging due to the intensity variations between the two recorded Raman spectra caused by unavoidable experimental changes as well as the presence of noise. Existent approaches suffer from drawbacks like spectral resolution loss, fluorescence residual, and artefacts. In this contribution, we proposed a reconstruction method based on non-negative least squares (NNLS), where the intensity variations between the two measurements are utilized in the reconstruction model. The method achieved fluorescence-free reconstruction on three real-world SERDS datasets without significant information loss. Thereafter, we quantified the performance of the reconstruction based on artificial datasets from four aspects: reconstructed spectral resolution, precision of reconstruction, signal-to-noise-ratio (SNR), and fluorescence residual. The artificial datasets were constructed with varied Raman to fluorescence intensity ratio (RFIR), SNR, full-width at half-maximum (FWHM), excitation wavelength shift, and fluorescence variation between the two spectra. It was demonstrated that the NNLS approach provides a faithful reconstruction without significantly changing the spectral resolution. Meanwhile, the reconstruction is almost robust to fluorescence variations between the two spectra. Last but not the least the SNR was improved after reconstruction for extremely noisy SERDS datasets.

1. Introduction

Raman spectroscopy is a label-free and non-destructive technology, which provides rich fingerprint information of almost all biomolecules and features a weak signal of water. Therefore, the technique is ideally suited to measure biological specimens and highly potential for *in-vivo* diagnostics [1–4]. The combination of chemometrics and Raman spectroscopy further improves the sensitivity, accuracy, and speed of Raman based detection [1,5]. All these benefits lead to the fast development of Raman spectroscopy in medical diagnostics [6–14], investigations of metabolism [15–17], intraoperative decision making [18,19], microbe identification [5,20,21] and many other biological investigations. However, biological samples often show a significant auto-fluorescence contribution [22], which is more intense compared to the Raman spectra and hampers further qualitative and quantitative analysis.

Up to date, instrumental and mathematical approaches have been proposed to overcome fluorescence background [23]. First, fluorescence can be suppressed via an excitation with near-infrared (NIR) laser sources because electronic transitions responsible for fluorescence are reduced by NIR excitation [24]. However, this requires longer integration time because Raman scattering decreases at a rate of λ^4.

---

[*] Corresponding author at: Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany; Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller University Jena, Helmoltzweg 4, 07743 Jena, Germany.
E-mail address: thomas.bocklitz@uni-jena.de (T. Bocklitz).

https://doi.org/10.1016/j.talanta.2018.04.050
Received 18 December 2017; Received in revised form 9 April 2018; Accepted 16 April 2018
Available online 22 April 2018
0039-9140/ © 2018 Elsevier B.V. All rights reserved.

Second, time-resolved Raman was reported to detect Raman scattering before fluorescence can take place [25]. Its wide application was yet hampered due to the high cost of short-pulse lasers. In the meantime, this technology is ineffective if the lifetime of fluorescence is comparable to the pulse length. Third, considering the different polarization properties of Raman scattering and fluorescence emission, polarization-resolved Raman was proven useful in gas-phase systems but fails for large molecules or matter in condensed phase [23,26]. On the other hand, mathematical baseline correction algorithms, like polynomial fitting [27], least squares [28], and extended multiplicative scatter correction (EMSC) [29], are widely applied due to their low cost and high flexibility. However, the parameters for the correction must be tuned specifically for different data, which is a crucial task. More importantly, if Raman bands are masked by an intense fluorescence background, the baseline correction might distort Raman peaks and mask important chemical information.

An alternative approach is shifted-excitation Raman difference spectroscopy (SERDS) [30–33], where two Raman spectra are measured with two slightly different excitation wavelengths. The shift between the two excitation wavelengths is kept small enough that the fluorescence remains unchanged but large enough that the Raman spectrum exhibits obvious shifts according to the excitation [31]. The fluorescence is removed in the difference of the two Raman spectra. However, the difference spectrum is hard to be interpreted because it does not directly show Raman bands. Meanwhile, the noise level of the difference spectrum is increased by a factor of $\sqrt{2}$ compared to the noise level of the two single spectra [34]. As shown in Fig. 1, the noise is higher in the difference spectrum (Fig. 1b) than within a single Raman spectrum (Fig. 1a) and the Raman bands are hardly visible in the difference spectrum. For this reason, spectral reconstruction is required to recover the fluorescence-free Raman spectrum, which can improve the SNR and its interpretability. Reconstructed examples are given in Fig. 1, from anti-derivative (antiD) method (Fig. 1c) and our new approach based on non-negative least squares (NNLS) (Fig. 1d). Both reconstruction methods are described in the subsection ‘spectral reconstruction’. Comparing to the difference spectrum, the noise in the reconstructed spectrum is apparently decreased and the Raman bands are clearly visible. It is also shown that the reconstructed SNR depends on the reconstruction approach. From Fig. 1, the anti-derivative method provided better SNR than the NNLS. Nonetheless, SNR is not the only criterion to evaluate a reconstruction approach, as will be present in this manuscript.

Already published reconstruction algorithms suffer from fluorescence residual, spectral resolution loss, or high-frequency artifacts [32,35–37]. Among those the fluorescence residual is a common issue of all reconstruction approaches. In fact, the intensities of the two recorded Raman spectra are often not identical due to experimental changes like fluctuations of laser power and photo-bleaching of molecules. The difference spectrum is hence not fluorescence-free, which contributes to fluorescence residual after reconstruction. A suitable intensity normalization could improve this issue [38], however, to a limited level. Therefore it is necessary to develop approaches capable of handling the undesirable intensity variations between measurements and allow to inspect the (estimated) fluorescence spectrum.

This contribution proposes a reconstruction procedure via non-negative least squares (NNLS), where we involved the experimental deduced intensity variations in the reconstruction model. The method was verified with three real-world SERDS datasets and we compared the results to the reconstruction by anti-derivative (antiD) and Fourier transform (FT) based methods. Thereafter, the performance of the reconstruction was quantified on the basis of artificial datasets from four aspects: reconstructed spectral resolution, precision of reconstruction, reconstructed signal to noise ratio (SNR), and fluorescence residual. The artificial SERDS datasets were constructed with varying values of spectral parameters including Raman to fluorescence intensity ratio (RFIR), SNR, full-width at half-maximum (FWHM), excitation wavelength shift, and fluorescence variation between the two spectra.

2. Material and methods

2.1. Experimental and Raman spectroscopy

The spectroscopy was based on different samples including 4-acetamedophenal and skin sample sectioned from pig ear. The measurement was performed using the excitation of a tunable laser source (785 ± 1 nm) with laser power of 20 mW. The Raman scattering was dispersed by a grating with 830 lines/mm and detected with a CCD camera (Andor iDus). Particularly, considering the largely different peak widths of Raman bands for 4-acetamedophenal, the measurement was split into two parts with different wavenumber regions: 640–1800 cm^{-1}, and 2600–3385 cm^{-1}. The three datasets are given in Fig. 2. In all three cases, the intensities varied significantly between the two recorded Raman spectra, resulted from the experimental changes. The influence of such variations on the spectral reconstruction will be revealed in the following context.

2.2. Simulation details

Besides the three real-world datasets, artificial SERDS datasets were constructed with varying FWHM, SNR, maximal Raman intensity ($r_{\max}$), excitation wavelength shift (m), and fluorescence variations between the two spectra. Each spectrum contained 1024 data points ($N = 1024$). The excitation wavelength shift represents the shift between Raman spectra measured with two excitation wavelengths counted in (spectral) data points. The fluorescence intensity was unchanged during the simulation while the intensity of Raman spectrum was changed by varying the maximal Raman intensity ($r_{\max}$). This equivalently changed the value of RFIR. For this reason, the two items, maximal Raman intensity and RFIR will be used interchangeably in this contribution.

The Raman spectrum was composed of five Gaussian peaks (Eq. (1a, b)) with various maximal intestines ($\vec{I}$) sharing the same standard deviation (σ_p). The standard deviation (σ_p) corresponds to

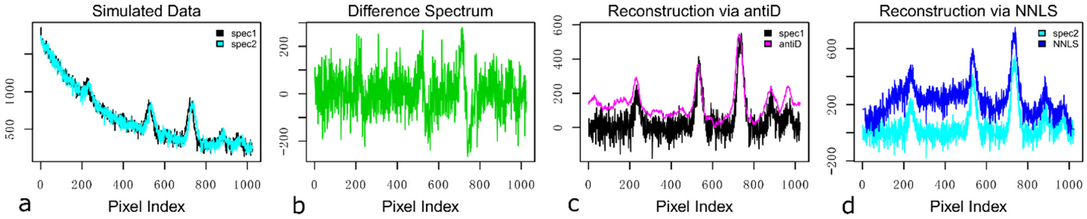


Fig. 1. The necessity of spectral reconstruction: without reconstruction, the difference spectrum (b) of two Raman spectra (a) is difficult to interpret and bears severer noise than the single measurement. Both interpretability and SNR were improved after reconstruction with antiD (c) and NNLS (d). The notation ‘spec1′ and ‘spec2′ refers to the first and second Raman spectrum in a SERDS dataset, respectively. The details of the antiD and NNLS are given in the following text.

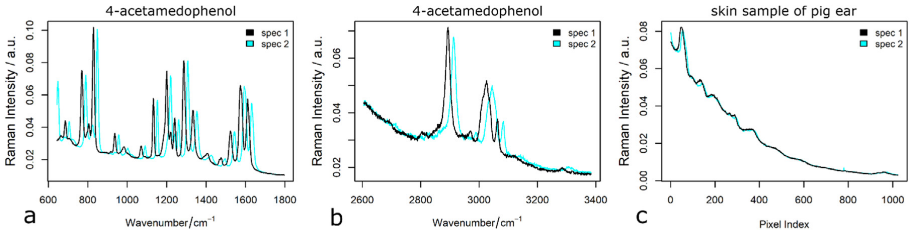


Fig. 2. Raw SERDS spectra measured from 4-acetamedophenal (a and b) and skin sample (c).

Table 1
investigated values of spectral parameters for simulation.

Parameters	Values
$r_{\max}$	250, 500, 1000
SNR	3, 4, 5, 6, 8, 10, 15, 20
σ_p	5, 10, 15, 20, 25, 30, 35
FWHM ($2\sqrt{2\ln(2)}\cdot\sigma_p$)	11, 23, 35, 47, 58, 70, 82
Excitation wavelength shift (m)	5, 8, 11, 14, 17, 20, 24, 26 29, 32, 35, 38, 41, 44, 47, 50

$\text{FWHM} = 2\sqrt{2\ln(2)}\cdot\sigma_p$. The fluorescence was simulated by an exponential function shown in Eq. (1c). To simulate the case where the fluorescence differs between the two spectra, $\overrightarrow{\boldsymbol{f}}^1$ and $\overrightarrow{\boldsymbol{f}}^2$ (Eq. (1e)) were used instead of $\overrightarrow{\boldsymbol{f}}$, with $p_f = 0.97$ (Eq. (1d)). More importantly, the simulation was performed with or without noise. For the former case, white noise with standard deviation of σ_n was generated separately for the two spectra. σ_n was determined by $\sigma_n = r_{\max}/SNR$ for a given SNR. Noteworthy, to overcome the randomness of generating noise, the simulation was repeated for 50 times by regenerating the noise using the same parameters. This led to 50 artificial datasets with the identical simulation parameters. The values of each parameter used for the simulation were summarized in Table 1. Fig. S1 visualizes example datasets constructed with $\sigma_p = 10$, $m = 11$, $\text{SNR} \in (3, 4, 5, 6, 8, 10, 15, 20)$, and $r_{max} \in (250, 500, 1000)$. As it was shown, the simulated datasets ranged from very poor quality to relatively good quality. This ensured us to obtain a faithful evaluation of the reconstruction.

$$
\begin{aligned}
&(a)\text{: } \overrightarrow{\boldsymbol{r}}^1 = r_{\max}\cdot\textstyle\sum_{i=1:5}\left(I_i\cdot\exp\left(\dfrac{(k+m-p_i)^2}{2\sigma_p^2}\right)\right)\\
&(b)\text{: } \overrightarrow{\boldsymbol{r}}^2 = r_{\max}\cdot\textstyle\sum_{i=1:5}\left(I_i\cdot\exp\left(\dfrac{(k-p_i)^2}{2\sigma_p^2}\right)\right)\\
&(c)\text{: } \overrightarrow{\boldsymbol{f}} = 1500\cdot\exp(-k/300) + 200\\
&(d)\text{: } a_k = \frac{1-p_f}{N/2-1}\cdot k + \frac{p_f\cdot N/2-1}{N/2-1}\\
&(e)\text{: } f_k^1 = f_k,\ f_k^2 = f_k/a_k\\
&(k = 1, 2, \ldots 1024,\ N = 1024)
\end{aligned}
\tag{1}
$$

3. Results and discussion

3.1. Spectral reconstruction

The existent spectral reconstruction approaches roughly include four categories: numerical peak fitting, anti-derivative, frequency manipulation, and shift matrix [36,39]. The numerical peak fitting suffers from the significant uncertainty especially for Raman bands overlapping with each other like in biological Raman-spectra. Hence only the other three strategies are investigated in this contribution. Anti-derivative is the most straightforward way to reconstruct a Raman spectrum. It is based on a one-dimensional integration for the difference spectrum, as shown in Eq. (2).

$$
\begin{aligned}
d_{k,m} &= s_{k+m} - s_k\\
r_k &= \sum_{i=1}^{k} d_{k,m}\cdot m
\end{aligned}
\tag{2}
$$

The Fourier transform based reconstruction is expressed in Eq. (3). Ideally, two Raman spectra of a SERDS dataset can be formulated as Eq. (3)(a-b), where m is the excitation wavelength shift counted by spectral data points (index). From mathematical point of view, a pure shift of a vector in spatial domain is equivalent to a multiplication with a phase factor $\exp(-i2\pi km/N)$ in the frequency domain. This corresponds to Eq. (3)(c-d) after the Fourier transform of the two spectra. The frequency component of a pure Raman spectrum is given in Eq. (3)(e), where the Raman spectrum $\overrightarrow{\boldsymbol{r}}$ can be recovered with an inverse Fourier transform. One of the major problems of Fourier transform based reconstruction is the high-frequency artefacts caused by the frequency leakage during the Fourier transform of a data series with limited length. The artefacts can be decreased by applying a cosine apodization after Fourier transform, but this sacrifices the reconstructed spectral resolution [35].

$$
\begin{aligned}
&(a)\text{: } s_k^1 = f_k + r_k\\
&(b)\text{: } s_k^2 = f_k + r_k(\rightarrow m)\\
&(c)\text{: } \tilde{s}_k^1 = \widetilde{f}_k + \tilde{r}_k\\
&(d)\text{: } \tilde{s}_k^2 = \widetilde{f}_k + \tilde{r}_k\cdot\exp(-i2\pi km/N)\\
&(e)\text{: } \tilde{r}_k = \frac{\tilde{s}_k^1 - \tilde{s}_k^2}{1-\exp(-i2\pi km/N)}
\end{aligned}
\tag{3}
$$

Unlike the frequency manipulation with Fourier transform, another approach is to formulate the excitation wavelength shift in original spectral domain. In principle, shifting of a vector $\overrightarrow{\boldsymbol{v}}^T$ by m points is equal to multiply the vector with a matrix: $\overrightarrow{\boldsymbol{v}}'^T = \mathbf{I}\times\overrightarrow{\boldsymbol{v}}^T$. Here, $\mathbf{I}$ is a binary matrix with ones at the semi-diagonal and zeros elsewhere. The offset between this semi-diagonal and the main diagonal is equal to m (as shown by $\mathbf{I}_r'$ in Fig. 3). The direction of the offset (lower or upper to the main diagonal) corresponds to the shift direction (right and left). In order to avoid the instability of the reconstruction caused by the singularity of $\mathbf{I}_r'$, we modified matrix $\mathbf{I}_r'$ by assigning the first m diagonal elements to ones. This is equivalent to keep the first m elements of the identity matrix unchanged when constructing the shift matrix $\mathbf{I}_r'$. On the other side, a measured Raman spectrum ($\overrightarrow{\boldsymbol{s}}$) is composed of pure Raman spectrum ($\overrightarrow{\boldsymbol{r}}$) and fluorescence emission ($\overrightarrow{\boldsymbol{f}}$), which can be mathematically described by a matrix multiplication: $\overrightarrow{\boldsymbol{s}}^T = [\mathbf{I}_r, \mathbf{I}_f]\times[\overrightarrow{\boldsymbol{r}}, \overrightarrow{\boldsymbol{f}}]^T$. $\mathbf{I}_r$ and $\mathbf{I}_f$ are identity matrices with dimension $N\times N$, for N data points within each spectrum. Consequently, the SERDS dataset can be modeled as Eq. (4), where $(\overrightarrow{\boldsymbol{s}}^1, \overrightarrow{\boldsymbol{s}}^2)^T$ is a column vector concatenated by the two recorded Raman spectra. Hereby the spectral reconstruction problem is transferred into a problem of inverse linear regression, where Eq. (4) should be solved for $\overrightarrow{\boldsymbol{r}}$ and $\overrightarrow{\boldsymbol{f}}$. This was achieved via least squares algorithm with non-negative constraints, i.e., non-negative least squares (NNLS). Noteworthy, the method

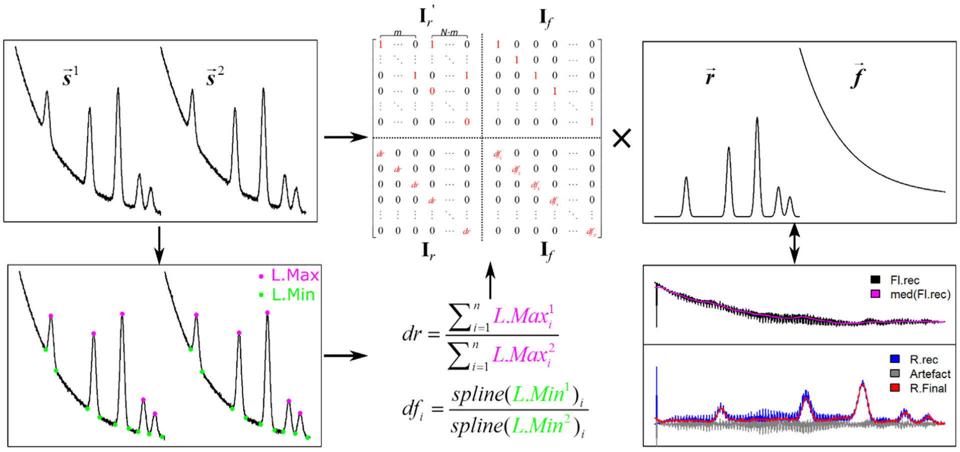

I_r'
I_f
m
N-m
$\vec{s}^1$
$\vec{s}^2$
I_r
I_f
×
$\vec{r}$
$\vec{f}$
L.Max
L.Min
$dr = \frac{\sum_{i=1}^{n} L.Max_i^1}{\sum_{i=1}^{n} L.Max_i^2}$
$df_i = \frac{spline(L.Min^1)_i}{spline(L.Min^2)_i}$
Fl.rec
med(Fl.rec)
R.rec
Artefact
R.Final

Fig. 3. Workflow of NNLS based spectral reconstruction: the identity matrices $\mathbf{I}_r$ and $\mathbf{I}_f$ are multiplied with regularization parameters representing the intensity variations between the two recorded Raman spectra. Specifically, the value *dr* was determined as the ratio of the sum of the local maxima of the two recorded spectra. Meanwhile, the fluorescence of both spectra was obtained by interpolating its local minima into the whole spectral region. The values of df_i $(i = 1, 2, \ldots, N)$ were calculated as the ratio between fluorescence intensities of the two spectra at each data point. Moreover, the artefacts after reconstruction is estimated by subtracting the reconstructed fluorescence signal (Fl.rec) with its smoothed signal (med(Fl.rec)) based on a strong median filter. Artefact-free Raman spectrum (R.Final) is obtained by adding the estimated artefacts (Artefact) to the reconstructed Raman signal (R.rec).

simultaneously recovers the pure Raman spectrum ($\vec{r}$) and fluorescence emission ($\vec{f}$).

$$(\vec{s}^1, \vec{s}^2)^T = \begin{bmatrix} \mathbf{I}_r' & \mathbf{I}_f \\ \mathbf{I}_r & \mathbf{I}_f \end{bmatrix} \times (\vec{r}, \vec{f})^T \tag{4}$$

The matrix shift based method is not new and has been reported in reference [40], which employed a series of measurements at more than two excitation wavelengths. Reconstruction from only two excitations would lead to significant fluorescence residual, as demonstrated in reference [39]. In fact, with Eq. (4), it was implicitly assumed that the spectral intensities remain the same between the two measurements, which rarely holds in practice. To take the intensity variations into consideration, we modified the matrices in Eq. (4) by introducing a scalar *dr* and a vector $\overrightarrow{df}$, as visualized in the upper row of Fig. 3. *dr* and $\overrightarrow{df}$ were calculated via the following steps. First, one spectrum was shifted spectrally to match the other one. Second, the local maxima and minima were located on the mean spectrum of the two matched spectra. Third, intensities of local maxima belonging to either spectrum were summed up. The intensity variation of the two Raman spectra was calculated as the ratio *dr* between the summed intensities. Meanwhile, the fluorescence of each spectrum was obtained by extrapolating its local minima into the whole spectral region. The df_i $(i = 1, 2, \ldots, 1024)$ was computed as the ratio between fluorescence signals at the corresponding data point.

Another problem of the NNLS method is the high-frequency artefacts caused by the computational approximate and the presence of noise. This is displayed in 'R.rec' and 'Fl.rec' in Fig. 3. Fortunately, the artefacts can be estimated from the reconstructed fluorescence ('Fl.rec'), providing the frequency of the fluorescence is extremely lower than that of the artefacts. Thereafter, the estimated artefacts were compensated in the reconstructed Raman spectra ('R.rec') to get an artefact-free Raman spectrum ('R.Final'). In our investigation, the artefacts were estimated by subtracting the reconstructed fluorescence with its smoothed signal by a strong median filtering (med(Fl.rec)). Noteworthy, the reconstruction is degraded at the boundary of the spectrum due to the modification of matrix $\mathbf{I}_r'$, as shown in Fig. 3 ('R.rec' and 'Fl.rec'), which can be corrected with the previously mentioned 'artefacts correction', as shown by Fig. 3 ('R.Final').

After introducing the spectral reconstruction approaches, it is important to evaluate their performance. This was accomplished based on real-world and artificial datasets. Qualitative evaluation was achieved with the real-world dataset, while the artificial datasets were applied for quantitative evaluation. In the latter case, quantitative markers of FWHM, SNR, and R^{12} [41] were calculated from the reconstructed spectrum. The Pearson's correlation coefficient (PCC, $\rho(X, Y) = \mathrm{cov}(X, Y)/(\sigma_X \cdot \sigma_Y)$) was computed between reconstructed and true Raman spectrum. Each of the markers evaluated the performance from a different aspect. The FWHM benchmarks the spectral resolution after reconstruction. The PCC represents the precision of the reconstruction with respect to the true spectrum. The SNR indicates the noise level after reconstruction. The R^{12} evaluates the fluorescence residual after the reconstruction, which was introduced to quantify the performance of computational baseline correction methods. A lower value of R^{12} represents a better baseline correction with less fluorescence residual but remaining Raman bands [41].

3.2. Qualitative comparison of reconstruction methods: Real-world dataset

The reconstructed results of the real-world datasets with the three approaches are given in Fig. 4. Fluorescence residual was obviously observed for antiD and FT based methods, due to intensity variations. The fluorescence residual was largely reduced and ignorable by NNLS reconstruction. The antiD method on the other side led to visibly broadened Raman bands or even completely lost bands highlighted by arrows in Fig. 4a. That means the spectral resolution was degraded by antiD method. In addition, the reconstruction by the FT based method was corrupted by high-frequency artefacts due to the frequency leakage [35]. Neither peak broadening nor the high-frequency artefacts were observed for NNLS approach, indicating the adequate performance of NNLS reconstruction.

Noteworthy, the reconstructed Raman peak positions are observed on different positions for the three reconstruction methods. The Raman peaks reconstructed with antiD lie in the middle of the corresponding peaks of the two recorded spectra. The FT and NNLS based approaches shift Raman bands to the positions of the Raman spectrum without

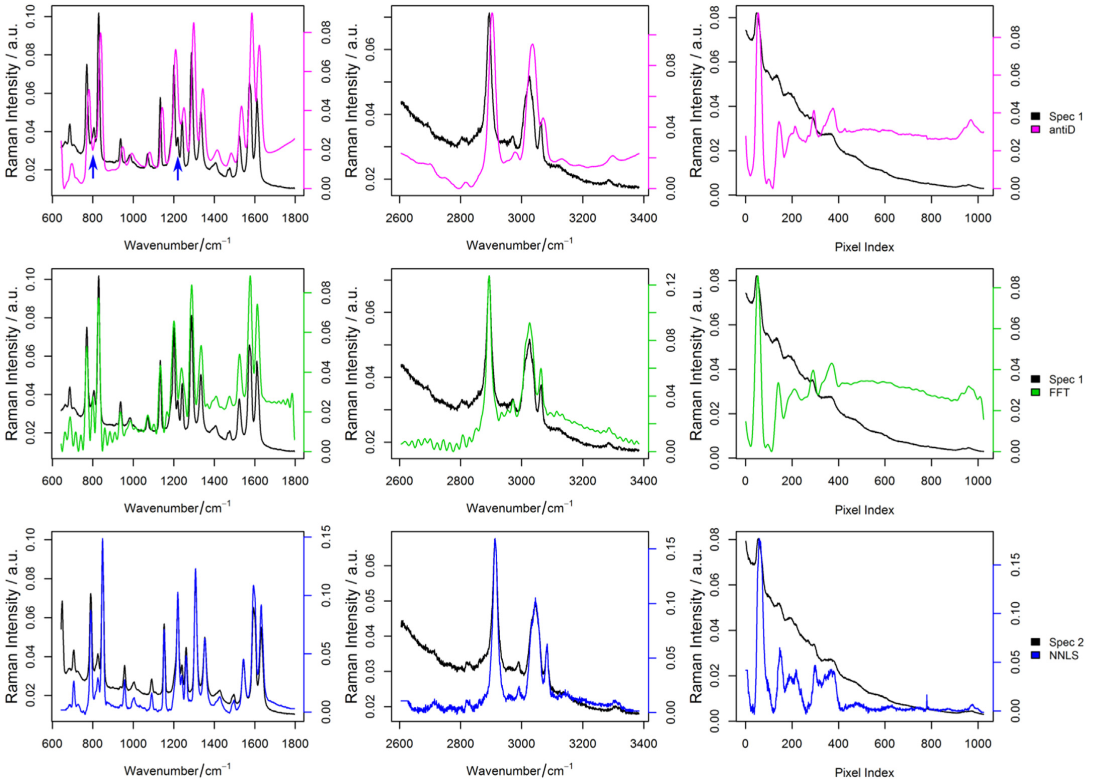


Fig. 4. Reconstructed spectra by three approaches. Obviously, FT and antiD based approaches failed to completely remove fluorescence. Meanwhile, the peaks were broadened or completely lost via antiD reconstruction (see blue arrows). On the other side, the FT based reconstruction was corrupted by severe high-frequency artefacts. NNLS gave the best reconstruction with ignorable remaining fluorescence and artefacts. Besides this, the peak widths were almost the same as the peak widths in the raw spectra after reconstruction and small peaks were also well reconstructed. Noteworthy, the reconstructed Raman peak positions were different for the three methods, dependent on which spectrum was treated as the spectrum without 'shift' during the calculation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

'shift', which in our implementation was the first and the second spectrum for FT and NNLS, respectively.

3.3. Quantitative comparison of reconstruction methods: Artificial datasets

The previous subsection gave a qualitative overview of the performance of the three reconstruction approaches. In this subsection a quantitative evaluation of the reconstruction methods is carried out with four markers based on artificial datasets. To do so, the markers including FWHM, PCC, SNR, and R^{12} were calculated after the reconstruction. The quantification of the FT reconstruction was not involved considering its inadequate performance caused by severe frequency leakage artefacts. In particular, the evaluation was based on two cases: noise-free and noise-containing simulation. The evaluation of spectral resolution (FWHM) and precision (PCC) was based on the noise-free artificial datasets in order to calculate FWHM and PCC more accurately without noise influence. In particular, the PCC and FWHM were calculated after an additional baseline correction on the reconstructed Raman spectrum in order to remove eventual influence caused by the fluorescence residual.

3.3.1. Noise-free artificial datasets: FWHM and PCC

The FWHM was calculated from the most intense Raman band. The other peaks were not used for the calculation because of two reasons. First, the five peaks within the same artificial spectrum possess the same FWHM. Second, the peaks with low intensities might be invisible after reconstruction and it is difficult to calculate the FWHM accurately based on these peaks. To calculate FWHM, the maximal Raman band in the reconstructed spectrum was fitted as a Gaussian profile. The FWHM was computed as $2\sqrt{2\ln(2)}\cdot\sigma_g$, where σ_g was the standard deviation of the fitted Gaussian profile. Thereafter, the ratio between the reconstructed FWHM and the value of FWHM used for the simulation was calculated to detect the change of the spectral resolution. The logarithm of the ratio was visualized in Fig. 5 relative to the values of FWHM and excitation wavelength shift used for the simulation. Each subplot provides a simulation with a different maximal Raman intensity (RFIR). Columns (a)-(c) and (e)-(f) correspond to results of NNLS and antiD, respectively. The first and second row corresponds to the simulation with identical and different fluorescence between the two spectra, respectively. The positive values, encoded as yellow and red, represent larger reconstructed FWHMs than the values used for simulation, i.e. a decreased spectral resolution. The negative values, encoded as blue, representing a smaller reconstructed FWHM than the true values, i.e., improved spectral resolution. The zeros, encoded as transparent revealed unchanged spectral resolution after the reconstruction.

As it is shown in Fig. 5, the reconstructed FWHM was increased with

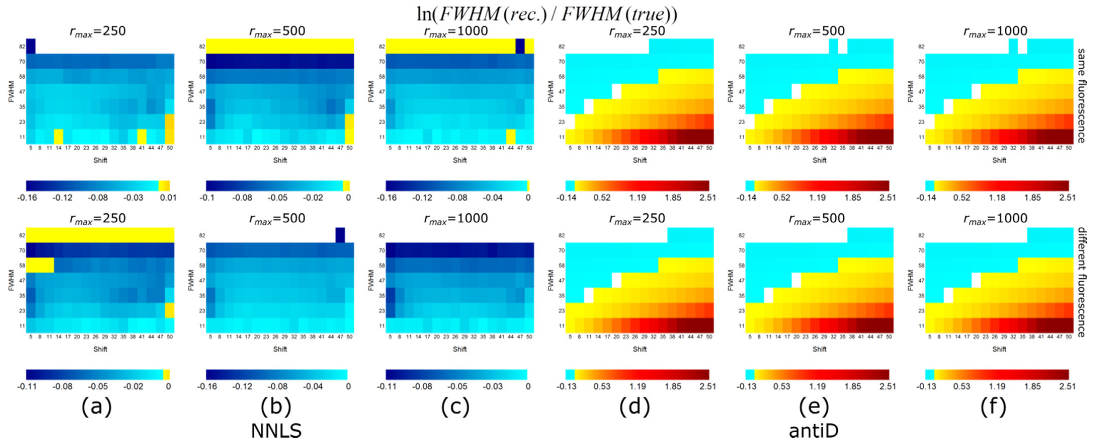


Fig. 5. The FWHM results of artificial datasets reconstructed with NNLS (a-c) and antiD (d-f) approach. The plotted values are the logarithm of the ratio between reconstructed and true FWHM. Each single plot visualizes the results from simulations with the same maximal Raman intensity (or RFIR) but different values of FWHM and excitation wavelength shift. The first and second row corresponds to the simulation with identical and different fluorescence, respectively. The positive values are encoded as yellow and red, while the negative values are encoded as blue. Apparently, the reconstructed FWHM was increased with antiD. In addition, the reconstructed FWHM was observed slightly affected by the variation of the fluorescence intensity, which can be seen by comparing the two rows. More importantly, the reconstructed FWHM was largely dependent on true FWHM and excitation wavelength shift. Nonetheless, the reconstructed FWHM was generally comparable to the true FWHM if a NNLS reconstruction was applied print in color.

antiD reconstruction especially for datasets with smaller FWHMs. That means the spectral resolution is decreased by antiD for these datasets. This is consistent with the results of the real-world datasets. Meanwhile, comparing the two rows of Fig. 5, the reconstructed FWHM was only slightly affected by the fluorescence variations. This was true for both approaches. If the maximal Raman intensity (RFIR) of the simulation was varied, the reconstructed FWHM changed for NNLS reconstruction but was almost stable for antiD reconstruction. On the other hand, the reconstructed FWHM was proven dependent on the FWHM and excitation wavelength shift, especially for antiD reconstruction. This can be explained by the theoretical background of SERDS technology. As it was discussed in ref [38], an unreasonable combination would lead to severe loss of spectral information. A reasonable combination of FWHM and excitation wavelength shift is thus an essential precondition for a usable SERDS measurement. Nonetheless, referring to the values shown in Fig. 5, the NNLS could reconstruct the Raman spectrum without significantly changing the spectral resolution as the ratio was generally very close to zero.

Besides the spectral resolution, the precision of the reconstruction was also verified. This was done by calculating the PCC between the reconstructed and the true spectra, which varies between zero and one. A higher PCC means a better precision. The results were visualized in Fig. S2, where the two rows correspond to simulations with identical and different fluorescence signals, respectively. The PCC was almost always above 0.8 for the NNLS reconstruction approach, demonstrating a satisfying reconstruction quality for different FWHMs and excitation wavelength shifts. This was different for the antiD reconstruction method, where the PCC was extremely low for data with a small FWHM due to the peak broadening after antiD reconstruction. Second, the PCC was dependent on the maximal Raman intensity, FWHM and excitation wavelength shift used for the simulation. The PCC increased for higher maximal Raman intensity (RFIR). An exception occurred for the antiD reconstruction on datasets simulated with identical fluorescence, where the PCC did not change with the maximal Raman intensity. Third, the PCC was dramatically dependent on true FWHM and excitation wavelength shift. Hereby, the antiD and NNLS behaved differently. The antiD performed better for larger FWHM and smaller excitation wavelength shift; while the NNLS gave better results for smaller FWHM and was less dependent on the excitation wavelength shift. To compare the results of the two rows shown in Fig. S2, the NNLS reconstruction is still slightly influenced by the fluorescence variations. This is because the simulated data contains very few Raman bands (local minima), leading to inadequate spline interpolation (see workflow in Fig. 3). Hence the values of df_i $(i = 1, 2, \ldots, N)$ for fluorescence variations correction are not accurate.

3.3.2. Noise-containing artificial datasets: SNR and R^{12}

Another important aspect for the performance evaluation is the reconstruction on noisy datasets. The details of the simulation can be found in the ‘material and methods’ section. Noteworthy, the excitation wavelength shift above 26 was not included in this subsection, because these values rarely occur in real-world measurements providing the wavelength shift between the two excitations is very slight. Figs. S3 and S4 visualize the reconstructed results based on NNLS and antiD procedures for datasets with $r_{\max} = 250$ and SNR = 3. Each sub-plot represents a different FWHM and excitation wavelength shift, involving 50 repetitions. The solid lines represent the average of the 50 reconstructed spectra, while the standard deviations over the 50 repetitions were plot as red shade. Obviously, the antiD was superior to NNLS by giving better reconstructed SNR. However, the fluorescence residual was smaller for NNLS than antiD. This fact is more obvious for data with higher SNR or higher r_{max}. The fluorescence residual for NNLS reconstruction originated in the inadequate spline interpolation as stated above. Nonetheless, the fluorescence residual is ignorable for datasets with large number of local minima like it is the case in real-world datasets (Fig. 4 (right)). In the following discussion, the performance of the reconstruction was evaluated with reconstructed SNR and R^{12}. These two markers were calculated for each repetition and the evaluation was based on the averaged values over the 50 repetitions.

As the first aspect, we compared the SNR of the reconstructed and the true Raman spectrum. To do so, we firstly estimated the noise level within the reconstructed spectrum by filtering the reconstructed spectrum with a S-G filter (p = 2, n = 51) and subtracting the result from the reconstructed spectrum. Thereafter, the reconstructed SNR was

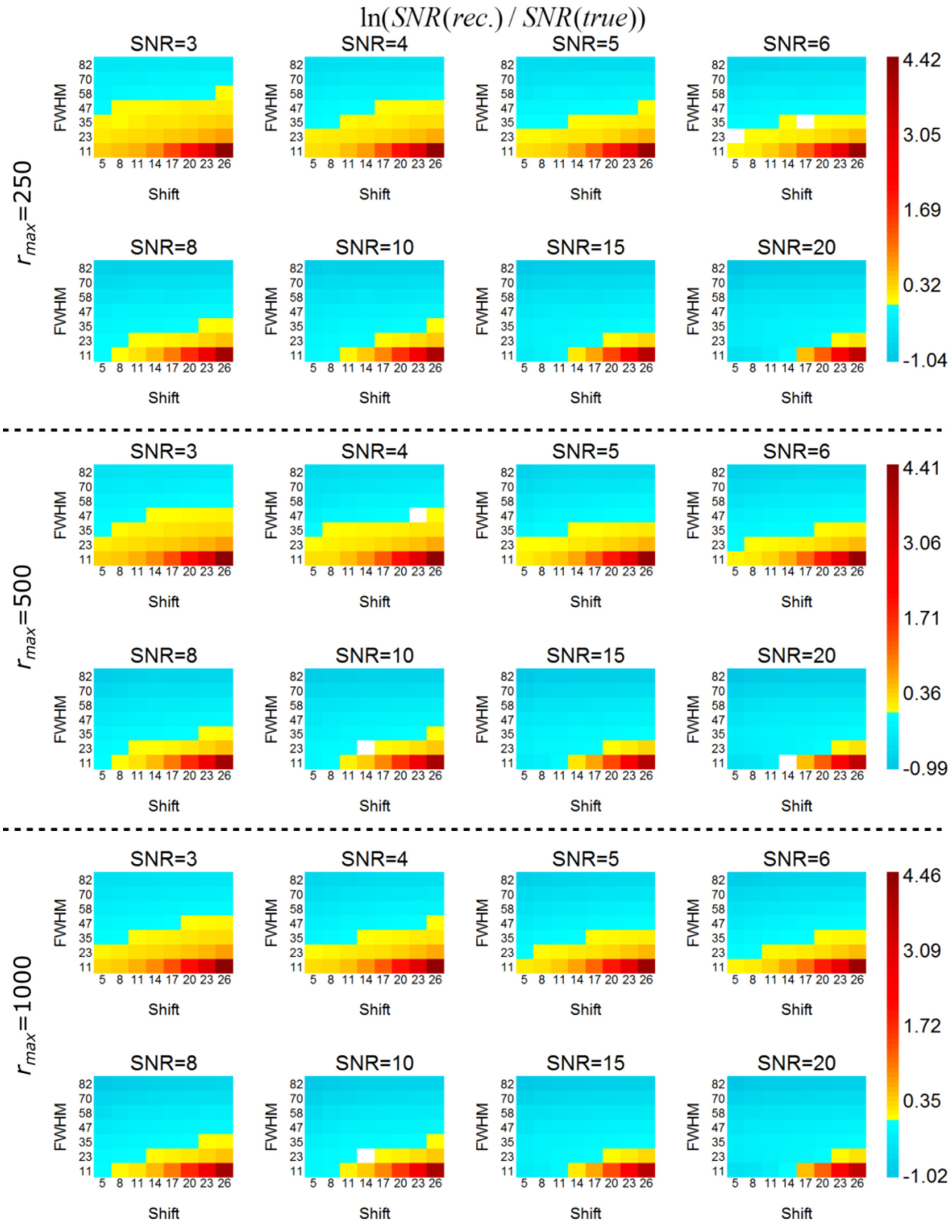


Fig. 6. The logarithm of the ratio between SNR of reconstructed and true Raman spectrum. The positive values are encoded as yellow and red, while the negative values are encoded as blue. With reasonable combinations of FWHM and excitation wavelength shifts, the reconstructed SNR could be comparable or equal to that of the single spectrum. For the datasets with extremely low SNR levels (SNR = 3, 4, 5, and 6), the reconstructed SNR was significantly improved comparing to the single spectrum. On the other hand, for cases of original SNR > 10, the reconstructed SNR demonstrated no further improvement. The reconstructed SNR was more likely to be improved than that of single spectrum for lower FRIR.

computed as the ratio between mean intensity of Raman peak region (highlighted in red color in Fig. S5) and the standard deviation of the estimated noise. Similarly, the SNR of the simulated Raman spectrum was calculated by $(\frac{1}{N}\sum_{i\in peaks} r_i)/\sigma_n$, where σ_n was the standard deviation of the simulated white noise. The logarithm of the ratio between reconstructed SNR and that of the single spectrum was plot in Fig. 6. The color code was similar as Fig. 5.

Clearly, the FWHM and excitation wavelength shift of the simulated spectra influenced the reconstructed SNR. With reasonable combinations, the reconstructed SNR could be comparable or equal to that of the single spectrum. In addition, the SNR level of the original single spectrum greatly influences the reconstructed SNR. For datasets with extremely low SNR levels (SNR = 3, 4, 5, and 6, for instance), the reconstructed SNR was significantly improved comparing to the single spectrum. On the other hand, for the single spectrum with large enough SNR, the reconstructed SNR exhibited no further improvement. Besides this, the value of maximal Raman intensity (RFIR) also matters for the reconstructed SNR. The reconstructed SNR was more likely to be increased for lower RFIR. Overall, it can be concluded that the SNR is possibly increased after reconstruction comparing to the single spectrum for extremely noisy datasets. However, according to the values shown in Fig. 6, the reconstructed SNR is not better than the mean spectrum of the two single Raman spectra.

A second aspect to check about reconstructions is the fluorescence residual. Ideally, the fluorescence is supposed to be completely removed from the reconstruct spectrum while all useful spectral information is retained. The verification of this point was done based on the quantitative marker R^{12} [41]. The regions with and without Raman bands were used for calculation of R^{12}, which were termed as peak and zero regions, respectively. The peak and zero regions used for calculating R^{12} were visualized in Fig. S6. According to the definition, a smaller R^{12} represents a better baseline correction, i.e., less fluorescence residual after reconstruction. Herewith we treated the original Raman spectrum as the perfect case where the fluorescence was completely removed, i.e., the ground truth. Any fluorescence residual after reconstruction will lead to larger reconstructed R^{12} than the ground truth. Accordingly, we calculated the logarithm of the ratio between the reconstructed and true R^{12} and visualized the results in Fig. S7.

The logarithm of the ratio was positive in general. That means the reconstruction could not completely remove the fluorescence. Also, the fluorescence residual was dependent on the parameters including FWHM, excitation wavelength shifts, RFIR of the simulated datasets. First, each sub-block in Fig. S7 displays simulations with a different SNR levels but the same RFIR. Apparently, the reconstruction was inferior for spectra with unreasonable combination of FWHM and excitation wavelength shift. Nonetheless, given a reasonable combination of FWHM and excitation wavelength shifts, the reconstruction was satisfying even for cases of low SNR. Second, comparing the results of datasets with the same SNR level but different RFIR, it was demonstrated that the reconstruction was superior for higher RFIR.

Besides the simulations using identical fluorescence contributions, the simulations containing fluorescence variations were carried out as well. The results were visualized in columns b, d, f of Fig. S8 and Fig. S9. Hereby the SNR was varied between 3, 5, and 10. The results from the identical fluorescence were re-plotted in columns a, c, and e for an easy comparison. Obviously, the reconstructed SNR was decreased while the R^{12} was increased comparing the results of identical fluorescence. For either marker, however, the changes were slight. That is to say, by involving fluorescence variations into the model, the NNLS reconstruction was almost tolerant to fluorescence variations.

3.4. Benchmark the reconstruction by the classification performance of an artificial task

Besides the quantitative analysis described above, we additionally benchmarked the performance of the SERDS reconstruction according to the classification performance. This was done based on a simulated binary classification task (Fig. S10 (a-c)). The details of the simulation and the results are included in the supporting information. Briefly, the dataset included five batches, each constructed with slightly different fluorescence. Meanwhile, the Raman intensity was varied slightly and randomly from spectrum to spectrum. The fluorescence of the two Raman spectra of each dataset varied according to Eq. (1). The mean spectrum of the difference spectra, the antiD reconstructed spectra and the NNLS reconstructed spectra are visualized in Fig. S10 (d-f), respectively. We constructed a partial least squares (PLS) classification and evaluated the classification performance based on a leave-one-batch-out cross-validation [42]. The mean sensitivity of the prediction was plotted in Fig. S11 relative to the number of latent variables used for PLS. The difference spectra and the antiD reconstruction produced the similar results. NNLS reconstruction led to much better prediction than the difference spectra and antiD reconstruction. That means the NNLS reconstruction enables to handle the fluorescence variations from batch to batch, even if the fluorescence was not completely removed after reconstruction.

4. Conclusion

This contribution revealed the limitations of SERDS reconstruction approaches based on antiD and Fourier transform, including fluorescence residual, spectral resolution loss, and severe artefacts. Thereafter, a spectral reconstruction approach based on NNLS was present. The intensity variations were estimated and modeled within the reconstruction, leading to a fluorescence-free reconstruction. Meanwhile, the high-frequency artefacts occurring via direct NNLS computation could be corrected without losing spectral resolution. The performance was benchmarked with three real-world SERDS datasets and a series of artificial datasets constructed with varying spectral parameters. It was demonstrated that the NNLS could provide almost fluorescence-free reconstruction without significantly changing the spectral resolution. In addition, SNR could be improved after reconstruction for extremely noisy raw SERDS datasets. The reconstructed SNR can be improved via antiD, because the noise contribution can be averaged during the integration. Moreover, the classification performance was better for NNLS reconstruction than antiD reconstruction.

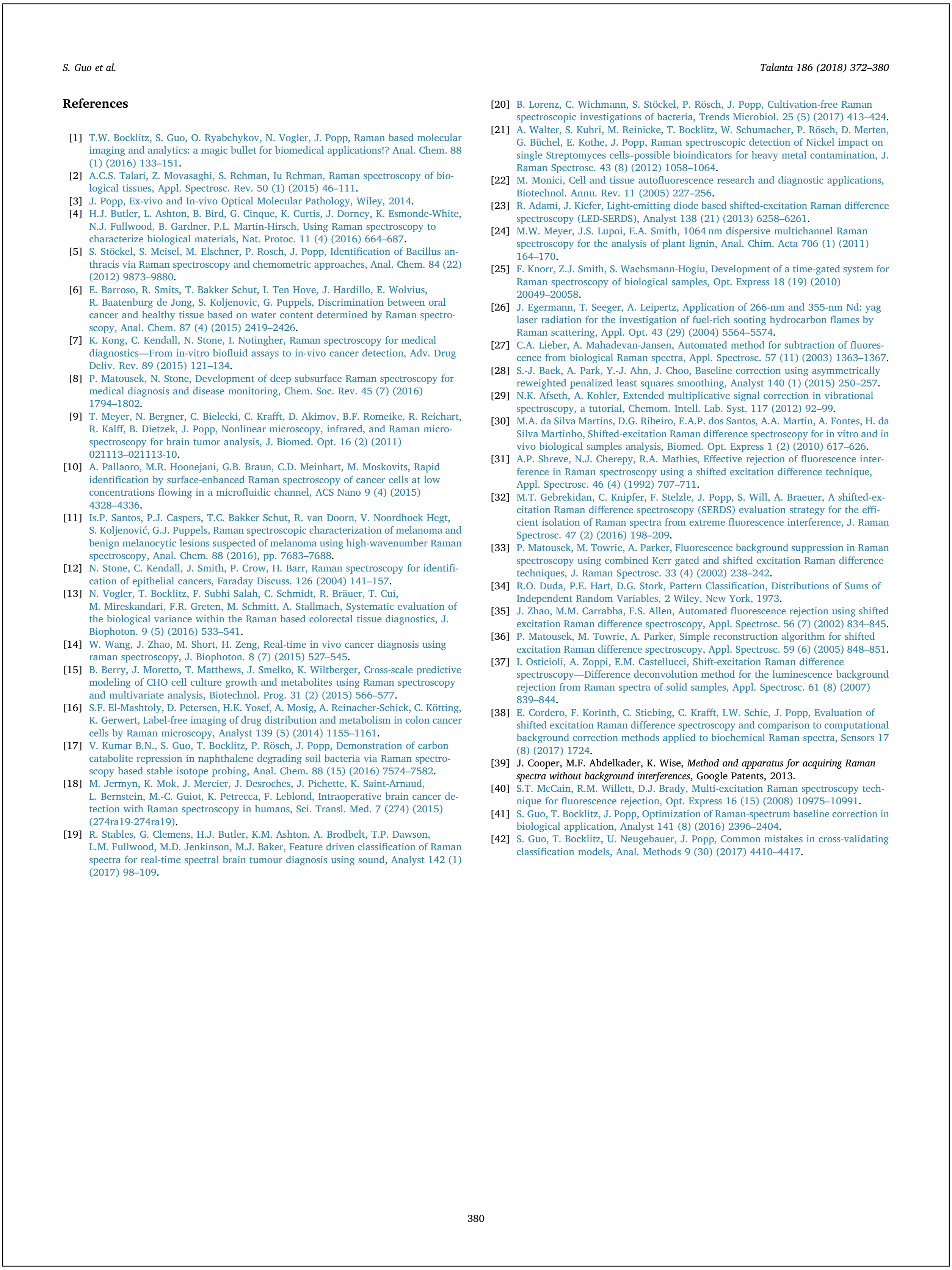
Nonetheless, the spectral parameters are important and limiting factors for the reconstruction methods. Especially the FWHM and excitation wavelength shift of the original dataset are important for the reconstruction methods. Particularly, for Raman bands with small FWHM, the NNLS approach is highly advantageous by providing unchanged spectral resolution and less fluorescence residual. In addition, the performance of NNLS is less dependent on the excitation wavelength shift compared to antiD reconstruction. On the other hand, for larger FWHM, antiD can lead to better PCC than NNLS. In addition, the reconstructed SNR can be improved via antiD by averaging the noise contributions during the integration. However, antiD degrades the spectral resolution obviously.

Acknowledgements

Financial supports of the BMBF and EU via the project ‘PhotoSkin’ (FKZ: 13N13243), the BMBF via the project “Intersept” (FKZ 13N13852), and the scholarship for S.G. from the ‘China Scholarship Council’ (CSC) are greatly appreciated.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.talanta.2018.04.050.



References

[1] T.W. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler, J. Popp, Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? Anal. Chem. 88 (1) (2016) 133–151.

[2] A.C.S. Talari, Z. Movasaghi, S. Rehman, Iu Rehman, Raman spectroscopy of biological tissues, Appl. Spectrosc. Rev. 50 (1) (2015) 46–111.

[3] J. Popp, Ex-vivo and In-vivo Optical Molecular Pathology, Wiley, 2014.

[4] H.J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N.J. Fullwood, B. Gardner, P.L. Martin-Hirsch, Using Raman spectroscopy to characterize biological materials, Nat. Protoc. 11 (4) (2016) 664–687.

[5] S. Stöckel, S. Meisel, M. Elschner, P. Rosch, J. Popp, Identification of Bacillus anthracis via Raman spectroscopy and chemometric approaches, Anal. Chem. 84 (22) (2012) 9873–9880.

[6] E. Barroso, R. Smits, T. Bakker Schut, I. Ten Hove, J. Hardillo, E. Wolvius, R. Baatenburg de Jong, S. Koljenovic, G. Puppels, Discrimination between oral cancer and healthy tissue based on water content determined by Raman spectroscopy, Anal. Chem. 87 (4) (2015) 2419–2426.

[7] K. Kong, C. Kendall, N. Stone, I. Notingher, Raman spectroscopy for medical diagnostics—From in-vitro biofluid assays to in-vivo cancer detection, Adv. Drug Deliv. Rev. 89 (2015) 121–134.

[8] P. Matousek, N. Stone, Development of deep subsurface Raman spectroscopy for medical diagnosis and disease monitoring, Chem. Soc. Rev. 45 (7) (2016) 1794–1802.

[9] T. Meyer, N. Bergner, C. Bielecki, C. Krafft, D. Akimov, B.F. Romeike, R. Reichart, R. Kalff, B. Dietzek, J. Popp, Nonlinear microscopy, infrared, and Raman microspectroscopy for brain tumor analysis, J. Biomed. Opt. 16 (2) (2011) 021113–021113-10.

[10] A. Pallaoro, M.R. Hoonejani, G.B. Braun, C.D. Meinhart, M. Moskovits, Rapid identification by surface-enhanced Raman spectroscopy of cancer cells at low concentrations flowing in a microfluidic channel, ACS Nano 9 (4) (2015) 4328–4336.

[11] Is.P. Santos, P.J. Caspers, T.C. Bakker Schut, R. van Doorn, V. Noordhoek Hegt, S. Koljenović, G.J. Puppels, Raman spectroscopic characterization of melanoma and benign melanocytic lesions suspected of melanoma using high-wavenumber Raman spectroscopy, Anal. Chem. 88 (2016), pp. 7683–7688.

[12] N. Stone, C. Kendall, J. Smith, P. Crow, H. Barr, Raman spectroscopy for identification of epithelial cancers, Faraday Discuss. 126 (2004) 141–157.

[13] N. Vogler, T. Bocklitz, F. Subhi Salah, C. Schmidt, R. Bräuer, T. Cui, M. Mireskandari, F.R. Greten, M. Schmitt, A. Stallmach, Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics, J. Biophoton. 9 (5) (2016) 533–541.

[14] W. Wang, J. Zhao, M. Short, H. Zeng, Real-time in vivo cancer diagnosis using raman spectroscopy, J. Biophoton. 8 (7) (2015) 527–545.

[15] B. Berry, J. Moretto, T. Matthews, J. Smelko, K. Wiltberger, Cross-scale predictive modeling of CHO cell culture growth and metabolites using Raman spectroscopy and multivariate analysis, Biotechnol. Prog. 31 (2) (2015) 566–577.

[16] S.F. El-Mashtoly, D. Petersen, H.K. Yosef, A. Mosig, A. Reinacher-Schick, C. Kötting, K. Gerwert, Label-free imaging of drug distribution and metabolism in colon cancer cells by Raman microscopy, Analyst 139 (5) (2014) 1155–1161.

[17] V. Kumar B.N., S. Guo, T. Bocklitz, P. Rösch, J. Popp, Demonstration of carbon catabolite repression in naphthalene degrading soil bacteria via Raman spectroscopy based stable isotope probing, Anal. Chem. 88 (15) (2016) 7574–7582.

[18] M. Jermyn, K. Mok, J. Mercier, J. Desroches, J. Pichette, K. Saint-Arnaud, L. Bernstein, M.-C. Guiot, K. Petrecca, F. Leblond, Intraoperative brain cancer detection with Raman spectroscopy in humans, Sci. Transl. Med. 7 (274) (2015) (274ra19-274ra19).

[19] R. Stables, G. Clemens, H.J. Butler, K.M. Ashton, A. Brodbelt, T.P. Dawson, L.M. Fullwood, M.D. Jenkinson, M.J. Baker, Feature driven classification of Raman spectra for real-time spectral brain tumour diagnosis using sound, Analyst 142 (1) (2017) 98–109.

[20] B. Lorenz, C. Wichmann, S. Stöckel, P. Rösch, J. Popp, Cultivation-free Raman spectroscopic investigations of bacteria, Trends Microbiol. 25 (5) (2017) 413–424.

[21] A. Walter, S. Kuhri, M. Reinicke, T. Bocklitz, W. Schumacher, P. Rösch, D. Merten, G. Büchel, E. Kothe, J. Popp, Raman spectroscopic detection of Nickel impact on single Streptomyces cells–possible bioindicators for heavy metal contamination, J. Raman Spectrosc. 43 (8) (2012) 1058–1064.

[22] M. Monici, Cell and tissue autofluorescence research and diagnostic applications, Biotechnol. Annu. Rev. 11 (2005) 227–256.

[23] R. Adami, J. Kiefer, Light-emitting diode based shifted-excitation Raman difference spectroscopy (LED-SERDS), Analyst 138 (21) (2013) 6258–6261.

[24] M.W. Meyer, J.S. Lupoi, E.A. Smith, 1064 nm dispersive multichannel Raman spectroscopy for the analysis of plant lignin, Anal. Chim. Acta 706 (1) (2011) 164–170.

[25] F. Knorr, Z.J. Smith, S. Wachsmann-Hogiu, Development of a time-gated system for Raman spectroscopy of biological samples, Opt. Express 18 (19) (2010) 20049–20058.

[26] J. Egermann, T. Seeger, A. Leipertz, Application of 266-nm and 355-nm Nd: yag laser radiation for the investigation of fuel-rich sooting hydrocarbon flames by Raman scattering, Appl. Opt. 43 (29) (2004) 5564–5574.

[27] C.A. Lieber, A. Mahadevan-Jansen, Automated method for subtraction of fluorescence from biological Raman spectra, Appl. Spectrosc. 57 (11) (2003) 1363–1367.

[28] S.-J. Baek, A. Park, Y.-J. Ahn, J. Choo, Baseline correction using asymmetrically reweighted penalized least squares smoothing, Analyst 140 (1) (2015) 250–257.

[29] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemom. Intell. Lab. Syst. 117 (2012) 92–99.

[30] M.A. da Silva Martins, D.G. Ribeiro, E.A.P. dos Santos, A.A. Martin, A. Fontes, H. da Silva Martinho, Shifted-excitation Raman difference spectroscopy for in vitro and in vivo biological samples analysis, Biomed. Opt. Express 1 (2) (2010) 617–626.

[31] A.P. Shreve, N.J. Cherepy, R.A. Mathies, Effective rejection of fluorescence interference in Raman spectroscopy using a shifted excitation difference technique, Appl. Spectrosc. 46 (4) (1992) 707–711.

[32] M.T. Gebrekidan, C. Knipfer, F. Stelzle, J. Popp, S. Will, A. Braeuer, A shifted-excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference, J. Raman Spectrosc. 47 (2) (2016) 198–209.

[33] P. Matousek, M. Towrie, A. Parker, Fluorescence background suppression in Raman spectroscopy using combined Kerr gated and shifted excitation Raman difference techniques, J. Raman Spectrosc. 33 (4) (2002) 238–242.

[34] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Distributions of Sums of Independent Random Variables, 2 Wiley, New York, 1973.

[35] J. Zhao, M.M. Carrabba, F.S. Allen, Automated fluorescence rejection using shifted excitation Raman difference spectroscopy, Appl. Spectrosc. 56 (7) (2002) 834–845.

[36] P. Matousek, M. Towrie, A. Parker, Simple reconstruction algorithm for shifted excitation Raman difference spectroscopy, Appl. Spectrosc. 59 (6) (2005) 848–851.

[37] I. Osticioli, A. Zoppi, E.M. Castellucci, Shift-excitation Raman difference spectroscopy—Difference deconvolution method for the luminescence background rejection from Raman spectra of solid samples, Appl. Spectrosc. 61 (8) (2007) 839–844.

[38] E. Cordero, F. Korinth, C. Stiebing, C. Krafft, I.W. Schie, J. Popp, Evaluation of shifted excitation Raman difference spectroscopy and comparison to computational background correction methods applied to biochemical Raman spectra, Sensors 17 (8) (2017) 1724.

[39] J. Cooper, M.F. Abdelkader, K. Wise, *Method and apparatus for acquiring Raman spectra without background interferences*, Google Patents, 2013.

[40] S.T. McCain, R.M. Willett, D.J. Brady, Multi-excitation Raman spectroscopy technique for fluorescence rejection, Opt. Express 16 (15) (2008) 10975–10991.

[41] S. Guo, T. Bocklitz, J. Popp, Optimization of Raman-spectrum baseline correction in biological application, Analyst 141 (8) (2016) 2396–2404.

[42] S. Guo, T. Bocklitz, U. Neugebauer, J. Popp, Common mistakes in cross-validating classification models, Anal. Methods 9 (30) (2017) 4410–4417.

Electronic Supplementary Material (ESI)

Spectral reconstruction for shifted-excitation Raman difference spectroscopy (SERDS)

Shuxia Guo[1,2], Olga Chernavskaia[1,2], Jürgen Popp[1,2,3], Thomas Bocklitz[1,2,*]
[1] Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany
[2] Institute of Physical Chemistry and Abbe Centre of Photonics, Friedrich-Schiller University Jena, Helmholtzweg 4, 07743 Jena Germany
[3] InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, 07743 Jena, Germany
* Email:Thomas.bocklitz@uni-jena.de

Abstract

Fluorescence emission has been one of the major obstacles to apply Raman spectroscopy in biological investigations. It is usually several orders more intense than Raman scattering and hampers further analysis. In cases where the fluorescence emission is too intense to be efficiently removed via routine mathematical baseline correction algorithms, an alternative approach is needed. One alternative approach is shifted-excitation Raman difference spectroscopy (SERDS), where two Raman spectra are recorded with two slightly different excitation wavelengths. Ideally, the fluorescence emission at the two excitations does not change while the Raman spectrum shifts according to the excitation wavelength. Hence the fluorescence is removed in the difference of the two recorded Raman spectra. For better interpretability a spectral reconstruction procedure is necessary to recover the fluorescence-free Raman spectrum. This is challenging due to the intensity variations between the two recorded Raman spectra caused by unavoidable experimental changes, as well as the presence of noise. Existent approaches suffer from drawbacks like spectral resolution loss, fluorescence residual, and artefacts. In this contribution, we proposed a reconstruction method based on non-negative least squares (NNLS), where the intensity variations between the two measurements are utilized in the reconstruction model. The method achieved fluorescence-free reconstruction on three real-world SERDS datasets without significant information loss. Thereafter, we quantified the performance of the reconstruction based on artificial datasets from four aspects: reconstructed spectral resolution, precision of reconstruction, signal-to-noise-ratio (SNR), and fluorescence residual. The artificial datasets were constructed with varied Raman to fluorescence intensity ratio (RFIR), SNR, full-width at half-maximum (FWHM), excitation wavelength shift, and fluorescence variation between the two spectra. It was demonstrated that the NNLS approach provides a faithful reconstruction without significantly changing the spectral resolution. Meanwhile, the reconstruction is almost robust to fluorescence variations between the two spectra. Last but not the least the SNR was improved after reconstruction for extremely noisy SERDS datasets.

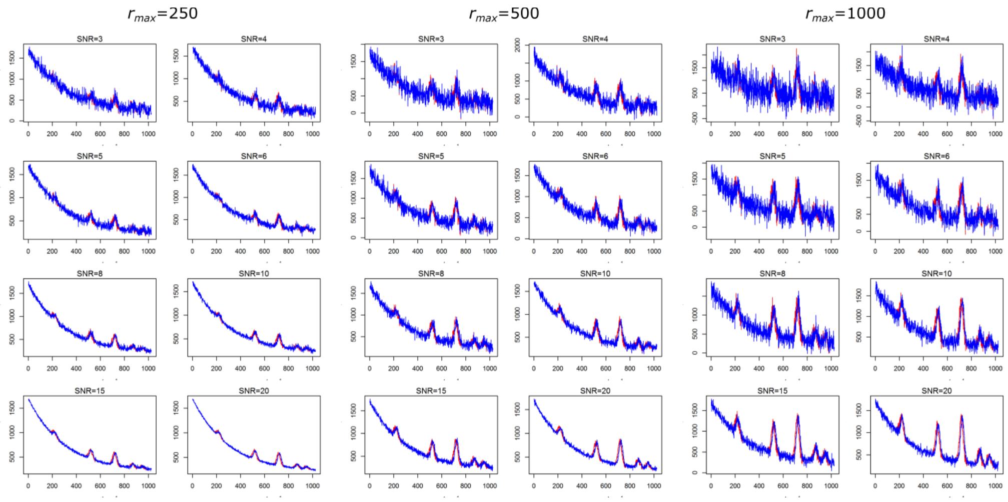


Figure S1 examples of artificial datasets simulated with parameters σ_p=10, m=11, $SNR \in (3,4,5,6,8,10,15,20)$, and $r_{max} \in (250,500,1000)$. the investigated values of SNR and MRI led to datasets from very poor quality to relatively good quality.

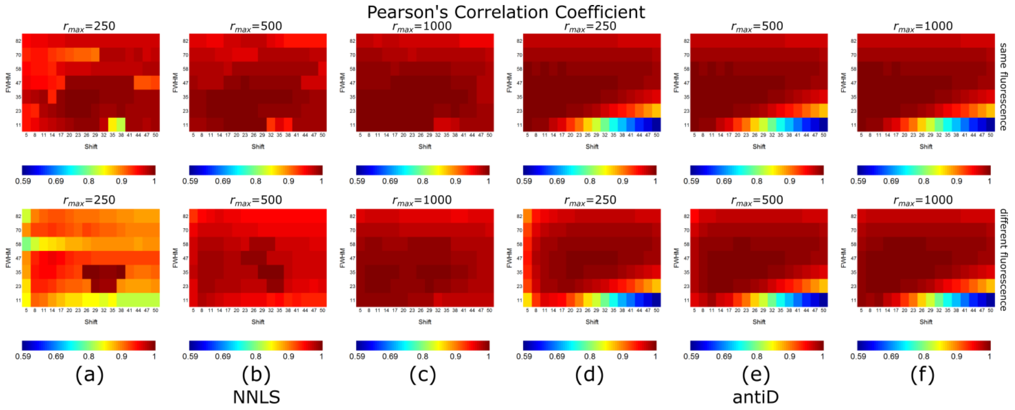


Figure S2: The results of PCC based on antiD and NNLS reconstruction methods. Apparently, the NNLS performs generally adequately for all cases of FWHM and excitation wavelength shift, which was not true for antiD. Higher maximal Raman intensity (or RFIR) led to larger PCCs for the same approach. However, the PCC did not change with maximal Raman intensity for antiD method, if the fluorescence is identical for the two spectra. Moreover, the PCC is dramatically dependent on FWHM and excitation wavelength shift. Reconstruction methods based on antiD leads to higher PCCs for larger FWHM and smaller excitation wavelength shift; while the PCC was higher for smaller FWHM and is less dependent on the excitation wavelength shift for NNLS.

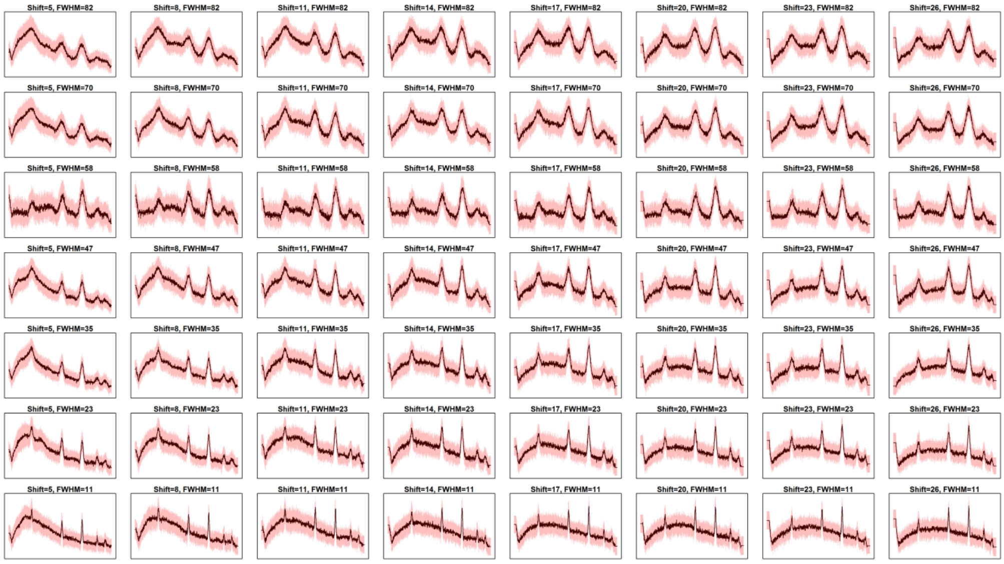


Figure S3 reconstructed results for datasets constructed with r_{max}=250, SNR=3. The white noise was regenerated 50 times for each value of excitation wavelength shift and FWHM. The black solid line represents the average of the 50 reconstructed spectra, while the red shade visualizes the variations over the 50 reconstructed spectra.

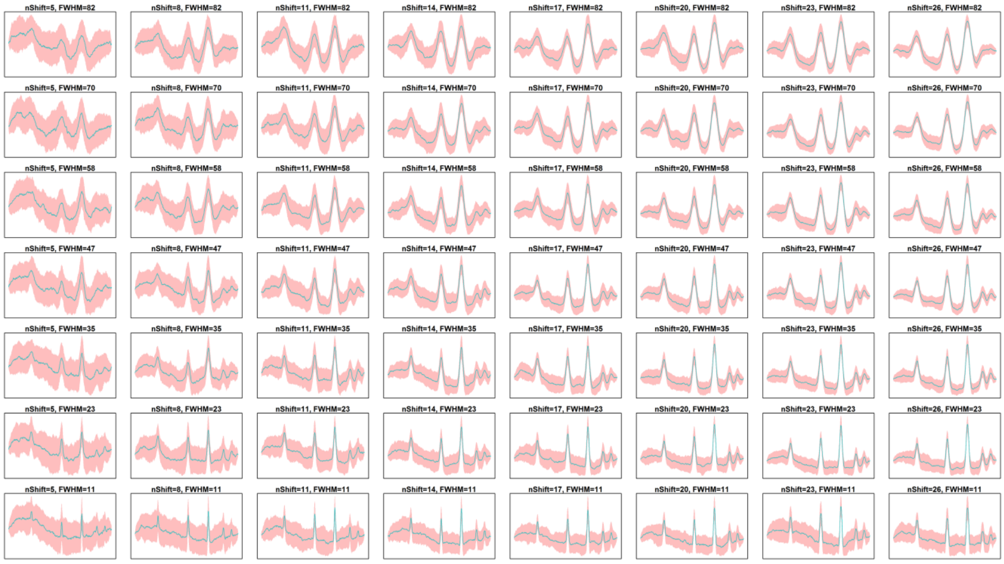


Figure S4 reconstructed results of antiD for datasets constructed with r_{max}=250, SNR=3, where the white noise was regenerated 50 times for each value of excitation wavelength shift and FWHM. The cyan solid line represents the average of the 50 reconstructed spectra, while the red shade visualizes the variations over the 50 reconstructed spectra.

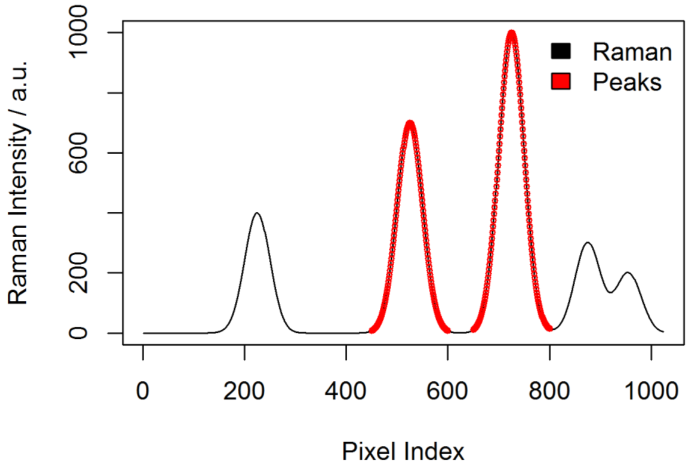


Figure S5 Peak positions used for calculating SNR

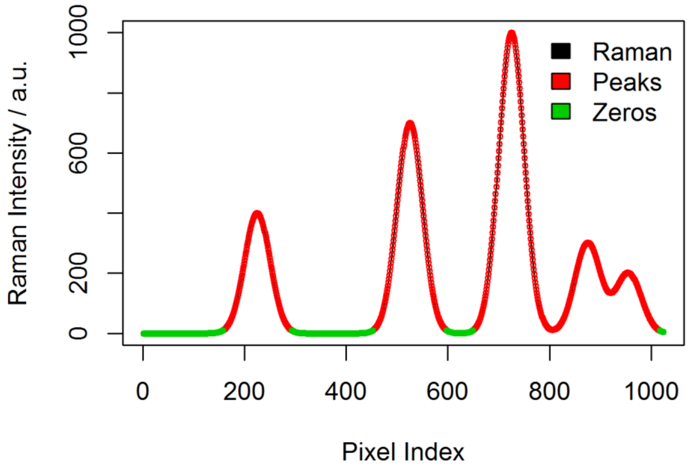


Figure S6 peak and zero regions used for calculating R^{12}.

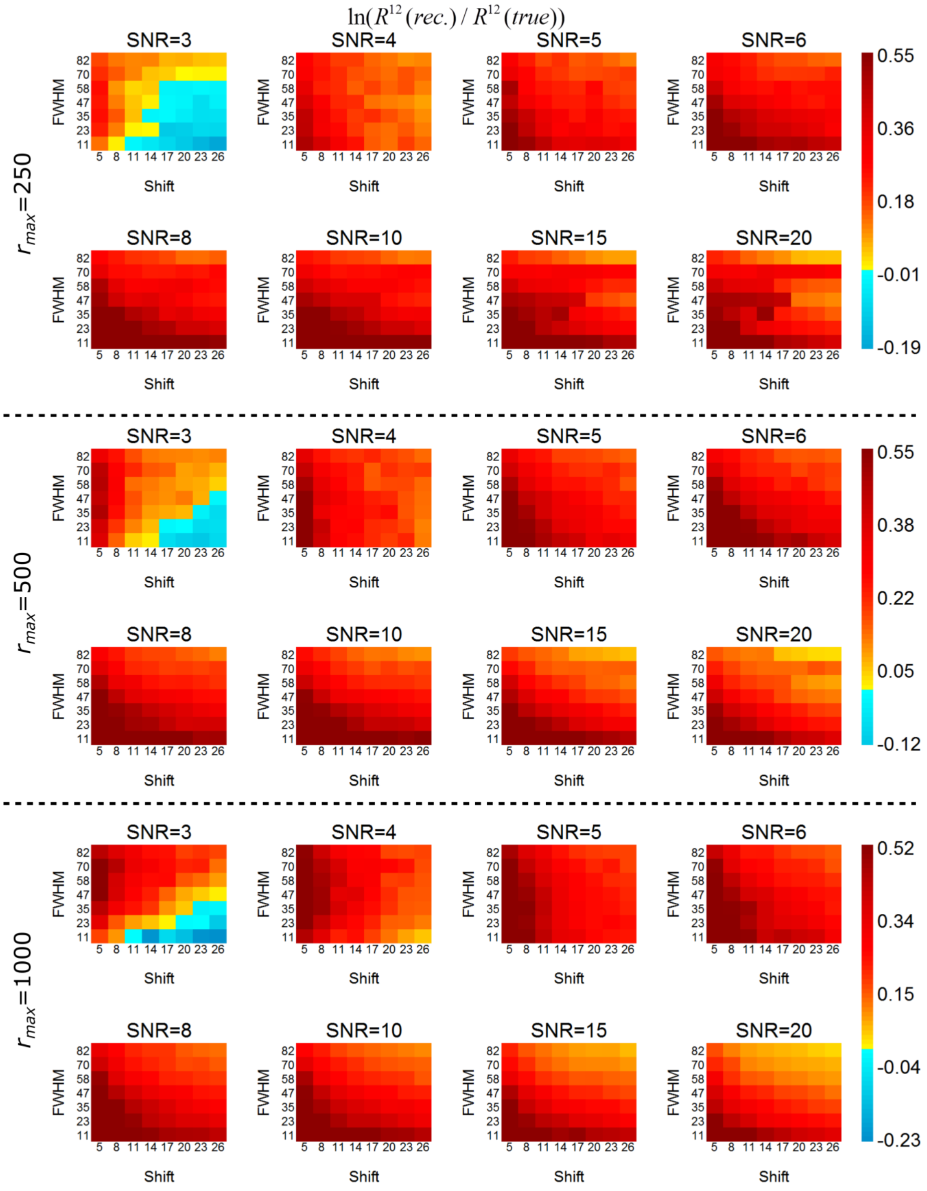


Figure S7: The logarithm of the ratio between the reconstructed and true R^{12}. Each sub-block separated by dash lines displays results of datasets with different SNR levels but the same maximal Raman intensity. The reconstruction was worse for spectra with higher noise levels, or unreasonable combination of FWHM and excitation wavelength shifts. However, for reasonable combinations of FWHM and excitation wavelength shifts, the reconstruction was satisfying even for extremely low SNRs. Further, the reconstruction was improved by a high RFIR.

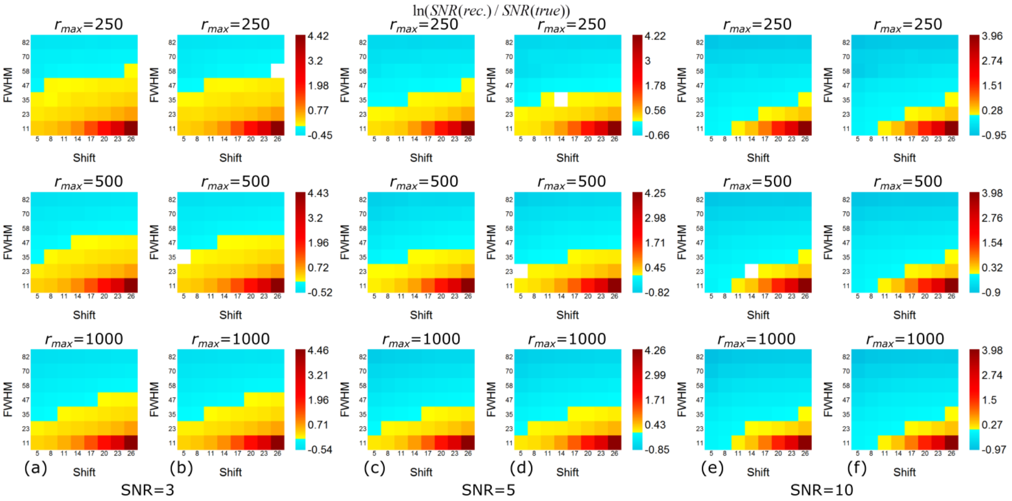


Figure S8 The SNR results of the reconstruction on noise-contained artificial datasets. The positive values are encoded as yellow and red, while the negative values are encoded as blue. Columns (a, c, e) correspond to the datasets constructed with identical fluorescence. The other columns refer to the datasets constructed with different fluorescence signals. The simulation was done for r_{max}=250, 500, or 1000, and SNR=3, 5, or 10. Obviously the reconstructed SNR was decreased due to the presence of fluorescence variation. However, the changes were rather slight. That is to say, the NNLS reconstruction was relatively tolerant to fluorescence variations.

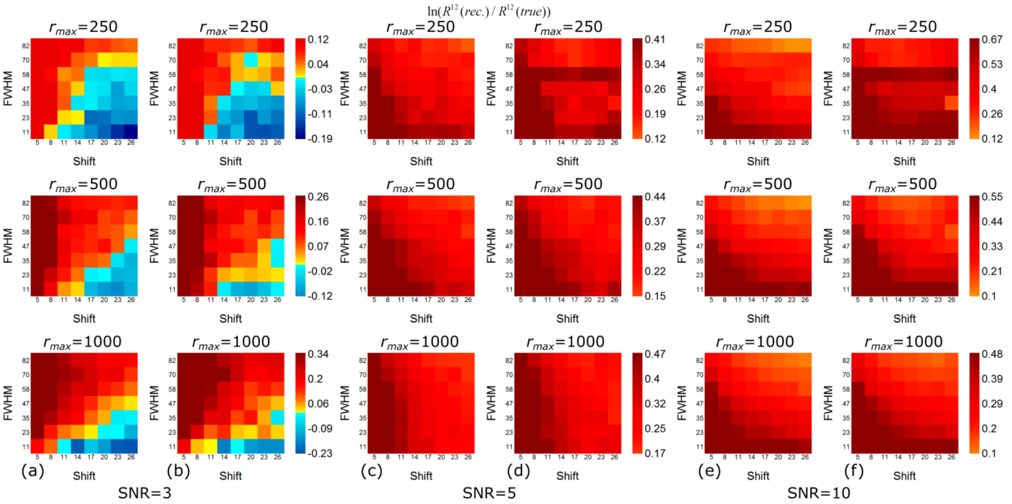


Figure S9 The R^{12} results of the reconstruction on noise-contained artificial datasets. The positive values are encoded as yellow and red, while the negative values are encoded as blue. Columns (a, c, e) correspond to the datasets constructed with identical fluorescence. The other columns refer to the datasets constructed with different fluorescence signals. The simulation was done for r_{max}=250, 500, or 1000, and SNR=3, 5, or 10. Obviously the reconstructed R^{12} was increased due to the presence of fluorescence variation. However, the changes were rather slight. That is to say, the NNLS reconstruction was relatively tolerant to fluorescence variations.

Simulation of binary classification

In order to check the performance of the SERDS reconstruction, we constructed a binary classification task based on artificial SERDS datasets constructed using the following parameters: r_{max}=250, SNR=3. The fluorescence for the two spectra of each SERDS dataset was varied according to Eq. (1). In particular, each group was composed of 100 spectra. The intensities of Raman bands slightly varied from spectrum to spectrum. Besides this, the 100 spectra were split into 5 batches and different fluorescence intensities were used for each batch. The mean spectra of the two groups are plotted in Figure S10 (a-c). The average of the difference spectrum and the reconstruction with antiD and NNLS methods for both groups were given in Figure S10 (d-f).

On this basis, we conducted a classification with partial least squares (PLS), where the number of the latent variables varied from 3 to 30. The classification performance was evaluated using a leave-one-batch-out cross-validation. The mean sensitivities corresponding to different reconstruction methods are shown in Figure S11. The classification results of the difference spectra and the antiD reconstruction were the same. The NNLS reconstruction led to much better prediction than the other two methods, despite the obvious fluorescence residual after the reconstruction.

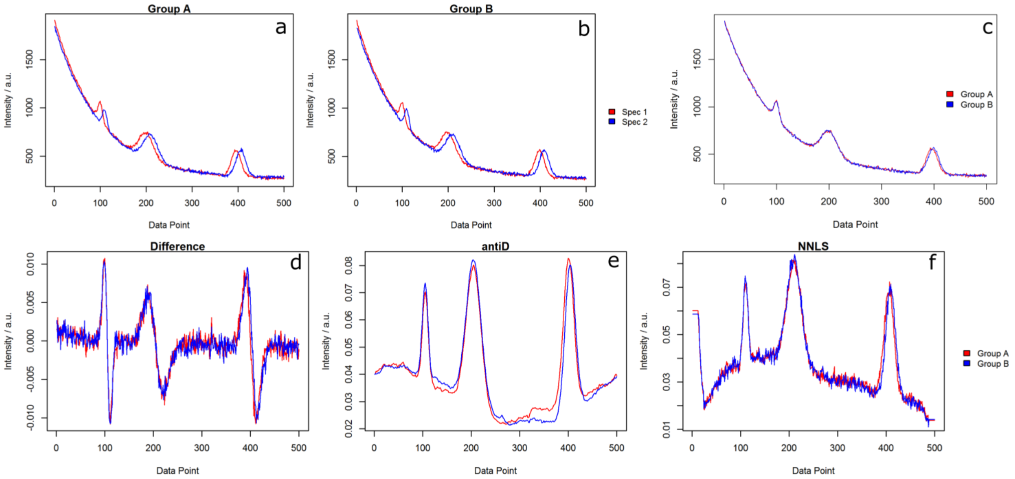


Figure S10 Mean spectra of the simulated data and the reconstructed results. (a-b) average of the two spectra of the SERDS dataset belonging to both groups. (c) mean spectrum of the two groups. (d) average of the difference spectrum for both groups. (e) average of the antiD reconstruction for both groups. (f) average of the NNLS reconstruction for both groups.

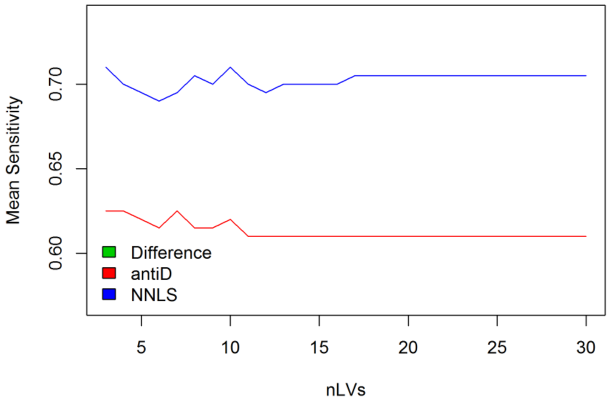


Figure S11 Mean sensitivities of the classification relative to the number of latent variables used for PLS. The results of difference spectra and antiD reconstruction were the same. The NNLS reconstruction led to much better prediction than the other two methods.

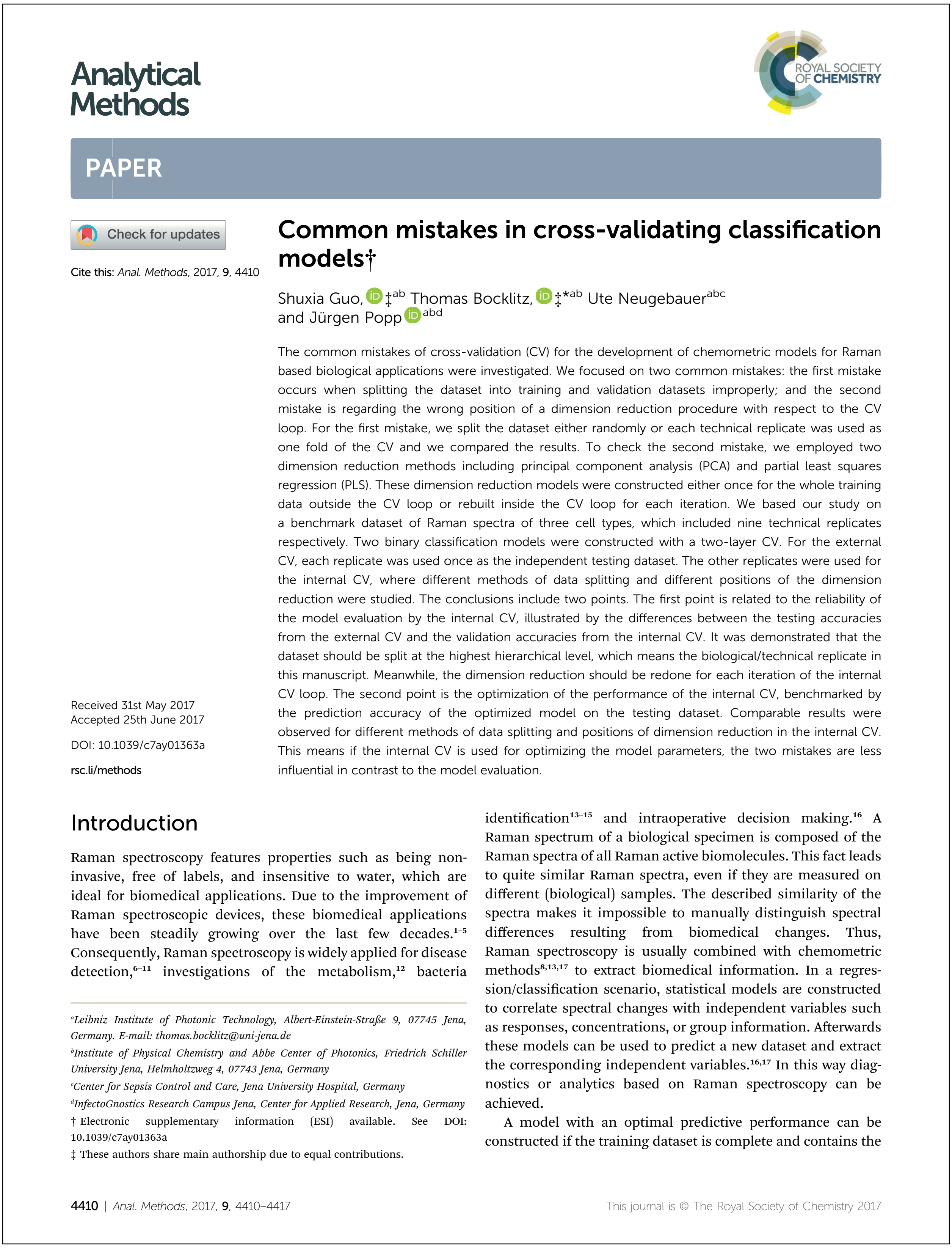
7.3 Common Mistakes in Cross-Validating Classification Models (A3)

S. Guo, T. Bocklitz, U. Neugebauer, and J. Popp, *Analytical Methods*, 2017, 9, 4410.

Der Nachdruck der folgenden Publikation erscheint mit freundlicher Genehmigung von Royal Society of Chemistry. Reprinted with kind permission from Royal Society of Chemistry.

Erklärungen zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation

Common mistakes in cross-validating classification models, S. Guo[1], T. Bocklitz[2], U. Neugebauer[3], and J. Popp[4], *Analytical Methods*, 2017, 9, 4410.				
Beteiligt an (*Zutreffendes ankreuzen*)				
	1	2	3	4
Konzeption des Forschungsansatzes	x	x		x
Planung der Untersuchungen	x	x		x
Datenerhebung			x	
Datenanalyse und -interpretation	x	x		
Schreiben des Manuskripts	x	x	x	x
Vorschlag Anrechnung Publikationsäquivalente	1.0			



Analytical Methods

PAPER

Check for updates

Cite this: *Anal. Methods*, 2017, **9**, 4410

Common mistakes in cross-validating classification models†

Shuxia Guo, ‡[ab] Thomas Bocklitz, ‡*[ab] Ute Neugebauer[abc]
and Jürgen Popp[abd]

The common mistakes of cross-validation (CV) for the development of chemometric models for Raman based biological applications were investigated. We focused on two common mistakes: the first mistake occurs when splitting the dataset into training and validation datasets improperly; and the second mistake is regarding the wrong position of a dimension reduction procedure with respect to the CV loop. For the first mistake, we split the dataset either randomly or each technical replicate was used as one fold of the CV and we compared the results. To check the second mistake, we employed two dimension reduction methods including principal component analysis (PCA) and partial least squares regression (PLS). These dimension reduction models were constructed either once for the whole training data outside the CV loop or rebuilt inside the CV loop for each iteration. We based our study on a benchmark dataset of Raman spectra of three cell types, which included nine technical replicates respectively. Two binary classification models were constructed with a two-layer CV. For the external CV, each replicate was used once as the independent testing dataset. The other replicates were used for the internal CV, where different methods of data splitting and different positions of the dimension reduction were studied. The conclusions include two points. The first point is related to the reliability of the model evaluation by the internal CV, illustrated by the differences between the testing accuracies from the external CV and the validation accuracies from the internal CV. It was demonstrated that the dataset should be split at the highest hierarchical level, which means the biological/technical replicate in this manuscript. Meanwhile, the dimension reduction should be redone for each iteration of the internal CV loop. The second point is the optimization of the performance of the internal CV, benchmarked by the prediction accuracy of the optimized model on the testing dataset. Comparable results were observed for different methods of data splitting and positions of dimension reduction in the internal CV. This means if the internal CV is used for optimizing the model parameters, the two mistakes are less influential in contrast to the model evaluation.

Received 31st May 2017
Accepted 25th June 2017

DOI: 10.1039/c7ay01363a

rsc.li/methods

Introduction

Raman spectroscopy features properties such as being non-invasive, free of labels, and insensitive to water, which are ideal for biomedical applications. Due to the improvement of Raman spectroscopic devices, these biomedical applications have been steadily growing over the last few decades.[1–5] Consequently, Raman spectroscopy is widely applied for disease detection,[6–11] investigations of the metabolism,[12] bacteria identification[13–15] and intraoperative decision making.[16] A Raman spectrum of a biological specimen is composed of the Raman spectra of all Raman active biomolecules. This fact leads to quite similar Raman spectra, even if they are measured on different (biological) samples. The described similarity of the spectra makes it impossible to manually distinguish spectral differences resulting from biomedical changes. Thus, Raman spectroscopy is usually combined with chemometric methods[8,13,17] to extract biomedical information. In a regression/classification scenario, statistical models are constructed to correlate spectral changes with independent variables such as responses, concentrations, or group information. Afterwards these models can be used to predict a new dataset and extract the corresponding independent variables.[16,17] In this way diagnostics or analytics based on Raman spectroscopy can be achieved.

A model with an optimal predictive performance can be constructed if the training dataset is complete and contains the

---

[a] *Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany. E-mail: thomas.bocklitz@uni-jena.de*

[b] *Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany*

[c] *Center for Sepsis Control and Care, Jena University Hospital, Germany*

[d] *InfectoGnostics Research Campus Jena, Center for Applied Research, Jena, Germany*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7ay01363a

‡ These authors share main authorship due to equal contributions.

This journal is © The Royal Society of Chemistry 2017

entire space of the population. However, in real applications this cannot be realized and the population space must be estimated by the sample space. Because of this fact a trade-off between the training error (bias) and the testing error (variance) exists. Thereby the testing error benchmarks the generalization performance of a model[18–20] and it is necessary to evaluate the aforementioned trade-off. A good solution for this task is to check the prediction of the model on an independent dataset. Unfortunately, it is expensive, time-consuming, or even impossible to generate enough independent data in most biological applications. Therefore, cross-validation (CV) has become a routine method to evaluate the models with limited data at hand. For CV, the available dataset is split into several parts, so called folds. Each fold is utilized as a test set once and predicted by the model developed on the basis of the other folds, which form the training data set.[21]

However, the CV is commonly performed with mistakes, which can lead to wrong results.[22] The first common mistake, which has a large influence on the result of a CV, is a wrong data splitting procedure. The methods for splitting the data have been widely investigated.[22–25] Based on the data splitting methods, a CV can be categorized into two types: exhaustive and non-exhaustive CVs. The methods belonging to the former type use all possible data splits and thus are computationally expensive. Therefore, they are less frequently applied, especially for large-size datasets.[21] The non-exhaustive CV is not complete because not all possible splits are tested. Hence, the results of this type of CV vary between different runs of the CV.[23,24] In this case, a proper data splitting is important, especially if significant variations exist within the groups of interest and between measurements. The results of a CV can be misleading, if the dataset is split improperly. This behaviour is demonstrated in Fig. 1, where a binary classification task is shown. The overall training dataset (black and red) includes two replicates for each group and an additional replicate acts as the independent testing dataset (blue). On the left side, the overall training dataset was split randomly into training and validation datasets (k-fold CV), while on the right, this splitting was done according to the replicate information (k-replicate CV). The two dotted lines represent the linear classification models, which ensure the largest margin between the two groups of the training dataset (black). For the k-fold CV, the model prediction on the validation dataset was perfect. In contrast, for the k-replicate CV, errors occur for the prediction on the validation dataset. If evaluated by the validation accuracy, the model on the left side is better. However, if the same classifiers are used to predict an independent test set, *e.g.* the real application case, the results will be reversed. This means that the classifier on the right side outperforms the other one with respect to the generalization performance. Therefore, the k-fold CV results in a misleading evaluation of the classifier on the left. In this contribution, this mistake is investigated based on real data; both the data splitting methods were applied and the results are compared.

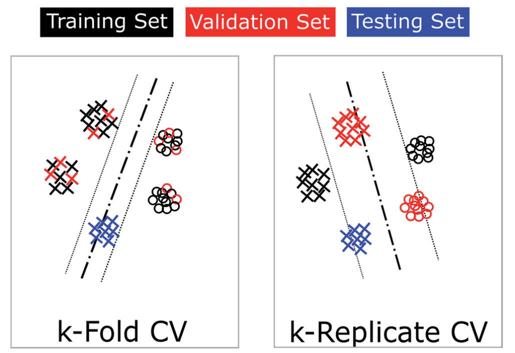


Fig. 1 Example of a binary classification task, with different data splitting methods. The overall training set (black and red) includes two replicates and an additional replicate acts as the independent testing dataset (blue). On the left, the overall training set was split randomly into the training and the validation datasets (k-fold CV), while on the right, this splitting was done according to the replicate information (k-replicate CV). The two separate planes represent the linear classifiers, which ensure the largest margin between the two groups in the training dataset (black) in both the cases.

The second common mistake is the improper position of dimension reduction with respect to the CV loop.[22] It was shown by Burden *et al.* that the pre-processing should be performed excluding the test set,[26] which means the respective test fold in the CV case. For this aspect, we utilized two dimension reduction methods: principal component analysis (PCA) and partial least squares regression (PLS). We compared the positions of the dimension reduction with respect to the CV loop. In one case we performed the dimension reduction once for the whole dataset outside the CV loop. In comparison, we applied the dimension reduction inside the CV loop and only used the actual training set to construct the dimension reduction model. These two CV types are termed outside-CV and inside-CV, respectively.

In this contribution we investigated both common mistakes and summarized our findings as guidelines to properly cross-validate classification models. We based our investigation on the Raman spectra of three cell types (MCF-7, BT-20, and OCI-AML3), which included nine replicates. Two binary classification models were developed and validated to separate MCF-7 from the other two cell types (BT-20 and OCI-AML3).

Experimental

Raman spectroscopy

The Raman spectroscopic measurements were described by Beleites *et al.*[27] and are briefly summarized in this section. The cells under investigation included two breast carcinoma derived tumour cells (MCF-7 and BT-20) and acute myeloid leukemia cells (OCI-AML3). Each cell type was cultivated five times and measured on nine days, which formed five biological replicates and nine technical replicates. We based our investigation on the nine technical replicates to have more folds for cross-validation with the k-replicate data splitting method. Raman spectra were measured with an excitation wavelength of 785 nm (xtra model, Toptica, Germany) and a power of 75 mW at the sample. An upright Raman microscope (Microprobe, Kaiser Optical Systems, USA) with a 60×/NA 1.0 water immersion objective (Nikon, Japan) was used. In total, 1553 cells (558 MCF-7, 477 BT-20, and 518 OCI-AML3) were measured in single-cell mode. The mean Raman spectra of the three cell types are plotted in Fig. S1.†

Computational

All computations were carried out in the statistical programming language Gnu R. The packages "signal",[28] "baseline",[29] "MASS"[30] and "e1071"[31] were utilized. The functions from these packages were complemented by in-house written procedures.

Data pre-processing

Before a classification model was constructed, the Raman spectra were pre-processed to remove spectral interferents and artefacts. The wavenumber axis was calibrated with a method described in ref. 32. The Raman band of CaF_2 at 322 cm^{-1} was used as a wavenumber standard. Afterwards, the wavenumber axis was interpolated on an equidistant grid with a step size of 1 cm^{-1}. Next the spectral background was removed by an asymmetric least squares baseline correction. The function 'baseline.als' within the R package 'baseline' was utilized and the parameters were set to lambda = 5 and $p = 0.001$. Thereafter, a 2-order Savitzky–Golay smoothing with a window width of 21 was applied. Finally, the wavenumber regions from 675 to 1785 cm^{-1} and from 2815 to 3020 cm^{-1} were used for classification after a vector normalization.

Models and validation

Classification. After the pre-processing, we developed two binary classification models: MCF-7 against BT-20 and MCF-7 against OCI-AML3. This was done by combining a dimension reduction technique with a classifier. The dimension was reduced by either principal component analysis (PCA) or partial least squares regression (PLS), without centring or scaling. The number of principal components (*n*PC), or latent variables (*n*LV) was varied from three to fifty. In particular, the PLS regression was applied using a dummy response variable (0 or 1) codifying the class belonging of the samples. To make the conclusions more general, we employed two methods for classification: a linear discriminant analysis (LDA) and a support vector machine (SVM) with linear kernel. A two-layer CV procedure was established as described below.

Cross-validation. A two-layer CV was used, as shown by the graphic workflow in Fig. 2 and the pseudo-code in Fig. S2.† The

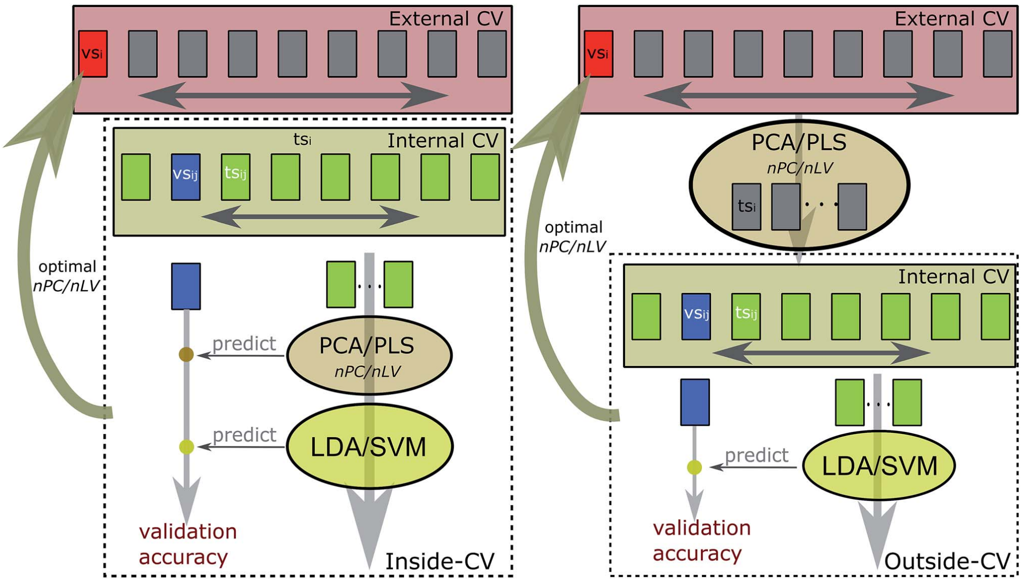


Fig. 2 Workflow of the applied two-layer CV. The classification model was a combination of a dimension reduction by PCA or PLS and a classifier (LDA or SVM). In detail, each replicate of the dataset was used once as the testing dataset (vs_i) within the external CV loop. The remaining replicates were used as the training dataset (ts_i) and split into the internal training dataset (ts_{ij}) and validation dataset (vs_{ij}) within the internal CV. The split was done either randomly or according to replicate information. In addition, the PCA or PLS model was built either only with ts_{ij} inside the internal CV loop (inside-CV) or once for the whole training dataset (ts_i) outside the internal CV loop (outside-CV). For each value of *n*PC (*n*LV), the model was rebuilt and validated by the internal CV, and the resulting eight validation accuracies were averaged. Thereafter, the optimal *n*PC/*n*LV featuring the highest average validation accuracy was used to build the model with the whole training dataset (ts_i) to predict the testing dataset (vs_i).

This journal is © The Royal Society of Chemistry 2017

internal CV was used to construct and validate the model, while the external CV was used to test the prediction performance of the model on independent data. For the external CV, each of the nine replicates was taken out once as an independent testing dataset (vs_i). The other eight replicates were used as the training datasets (ts_i) and an internal CV was carried out on the training datasets. In detail, ts_i was split into eight folds; each fold was used as a validation dataset (vs_{ij}) and predicted once. The resulting eight accuracies were averaged and used as validation results. The internal CV was repeated for each value of nPC (nLV). Afterwards, nPC (nLV) featuring the highest average validation accuracy of the internal-CV was used to construct a statistical model based on the overall training dataset (ts_i). This model was used for predicting the corresponding independent testing dataset (vs_i) and the testing accuracy was calculated.

The internal CV was carried out in four different ways. First, the training dataset (ts_i) was split either randomly into 8 folds or each biological replicate was used as one fold. Secondly, the position of the dimension reduction was varied. The PCA or PLS was performed either inside the internal CV or outside the internal CV but inside the external CV (see Fig. 2). For the inside-CV, the PCA (PLS) was rebuilt for each iteration of the internal CV loop based on the internal training dataset (ts_{ij}). The validation dataset (vs_{ij}) was predicted by this model. For the outside-CV, the PCA (PLS) model was built based on the overall training dataset (ts_i). Then the scores were split into the internal training dataset and validation dataset for the internal CV.

Results and discussion

Dataset characteristics

Before the discussion of the results can be done, it is necessary to describe the samples and the sampling hierarchy. We investigated 1553 single cell Raman spectra (558 MCF-7, 477 BT-20, and 518 OCI-AML3). These three cell types were derived from three cell lines and cultivated in five (biological) replicates and measured on nine days. Besides the differences between cell types, there exist significant variances between the replicates. This makes the dataset ideally suited to investigate the influence of the dataset splitting mechanisms used in the CV. In addition, MCF-7 and BT-20 are breast carcinoma cells and OCI-AML3 is acute myeloid leukemia cells, which means that they feature different similarities. Accordingly, the binary task MCF-7 against BT-20 is more difficult than the task of MCF-7 against OCI-AML3. This fact allowed us to study two classification tasks with different difficulty levels.

Performance

With the help of this complex dataset, we compared k-replicate CV with k-fold CV. The position of the dimension reduction (inside-CV and outside-CV) was studied as well, of which the workflow and the pseudo code are shown in Fig. 2 and S2,† respectively. In most applications, the validation accuracies are used for evaluating the performance of the model. Therefore, we plotted the validation accuracies from the first iteration of the external CV against nPC (nLV) in Fig. 3 as an example. Hereby the LDA was utilized for classification. The results of the SVM are shown in Fig. S3.† Apparently, the validation accuracies varied among different data splitting methods and positions of the dimension reduction in the internal CV loop. That is to say, these two aspects play an important role in the evaluation of the model. To get an idea of the reliability of the model evaluation, we performed an external CV for each case of internal CV. The independent testing dataset was predicted with the model giving the highest average validation accuracy. The testing and validation accuracies were compared. The idea behind this is that the testing and the validation accuracy are supposed to be consistent if the model is reliably evaluated. We visualized the testing accuracies (Test Acc) with LDA and a SVM as classifiers with black boxes in Fig. 4 and S4,† respectively. Meanwhile the highest average validation accuracies (Val Acc) were plotted as grey boxes. The applied data splitting methods, dimension reduction methods and their position for the internal CV are given as x-axis labels, where 'R' and 'F' represent the k-replicate CV and k-fold CV, respectively. The labels 'I' and 'O' denote the inside-CV and outside-CV, respectively. The marked p-values were obtained from a Wilcoxon-test, which is described in the following text.

A first aspect to discuss is the validity of the internal CV in the case of model evaluation. The most important criterion for this aspect is that the validation accuracies should be consistent with the testing accuracies. According to Fig. 3, the average validation accuracies were very high, if the dimension reduction was applied outside the internal CV (outside-CV). However, the values decreased for the inside-CV. This holds true for both data splitting methods. To understand this phenomenon, we need to refer to Fig. 4, where we compared the validation and testing accuracies. For the inside-CV, the two types of accuracies coincided, indicating that the validation accuracies fairly evaluated the model. In contrast, for the outside-CV, the testing accuracies were much lower than the validation accuracies. This means that the validation accuracies from the outside-CV were too high and the model was over-estimated. On the other hand the behaviours of PCA and PLS were different. In Fig. 3 the deviations of the validation accuracies between outside-CV and inside-CV are much larger for PLS than for PCA. This behaviour resulted from the different theoretical backgrounds of the two dimension reduction methods. As per PLS, a supervised method, the projection vectors are optimized according to the class labels to ensure the maximal separability among different classes. In the case of the outside-CV with PLS, the optimized projection vectors include the class label information of both the training and validation datasets. Therefore, the model is seriously over-estimated. To avoid such mistakes, the PLS should always be performed inside the internal CV. This is different for PCA, an unsupervised method. The eigenvectors of the PCA are extracted excluding the class information of the samples to capture the main sources of variability within the dataset. Thus the model evaluation for PCA is more robust to its mistaken position in the CV loop compared to PLS. However, there is still a possibility of over-estimation for PCA with outside-CV. In the outside-CV the PCA is performed on more

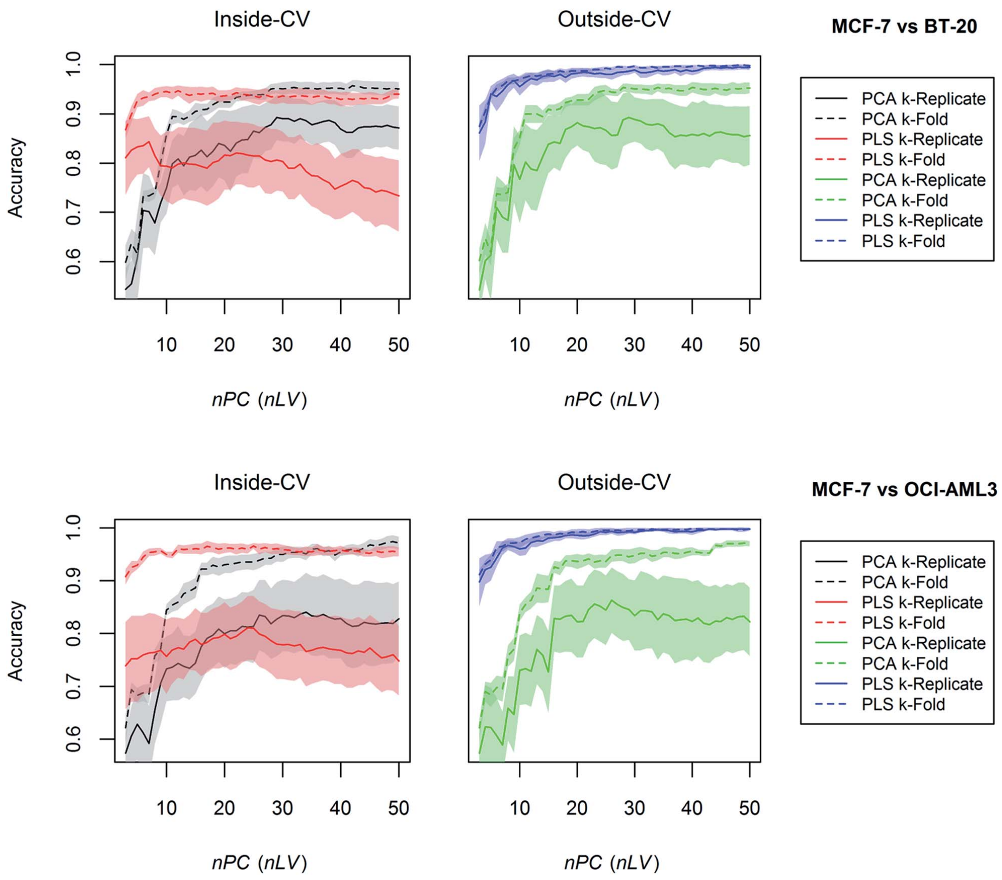


Fig. 3 Validation accuracies of the internal CV for the first iteration of the external CV. The two binary classification tasks (MCF-7 *vs.* BT-20 (*m vs. b*) and MCF-7 *vs.* OCI-AML3 (*m vs. o*)) were achieved by LDA. In both tasks, the outside-CV yields higher accuracies than the inside-CV. This is more obvious if supervised dimension reduction methods, such as PLS, are applied. On the other hand, the validation accuracies are always higher if a k-fold CV is used compared with a k-replicate CV. This effect is more enhanced for supervised dimension reduction.

samples, and thus the resulting eigenvectors tend to be closer to the population (true) eigenvectors.[33] This leads to higher validation accuracies compared to the values of an inside-CV. However, the improvement depends on how complete the training dataset represents the population. In an extreme case, if the training dataset is complete, the prediction of the testing dataset would not be improved by including the testing dataset into the PCA model. This means that the outside-CV with PCA does not necessarily lead to higher validation accuracies, *i.e.*, an over-estimation. However, the result of the outside-CV may vary significantly for different validation datasets, since the validation dataset is involved in the model construction. Especially, a serious over-estimation is more likely to occur, if the validation datasets bear significant variances compared with the training dataset. Above all, an inside-CV is highly recommended for the internal CV as a model evaluation approach.

A second aspect to discuss is the influence of the data splitting on the validity of the internal CV in the model evaluation case. According to Fig. 3, the k-fold CV always yields higher validation accuracies than the k-replicate CV, no matter which or where the dimension reduction method was performed. This is because an over-estimation occurs with k-fold CV, as demonstrated in Fig. 4, where the testing accuracies are lower than the validation accuracies for k-fold CV. In contrast, a k-replicate CV gave reliable model evaluation, demonstrated by the consistence of the two types of accuracies as shown in Fig. 4. The reason for this behaviour is that in a k-fold CV the dataset was split randomly without considering the hierarchical properties of the dataset. The information within the validation dataset has been implicitly used during the model construction. The independence condition for the training and the validation datasets was violated. Moreover, the

This journal is © The Royal Society of Chemistry 2017

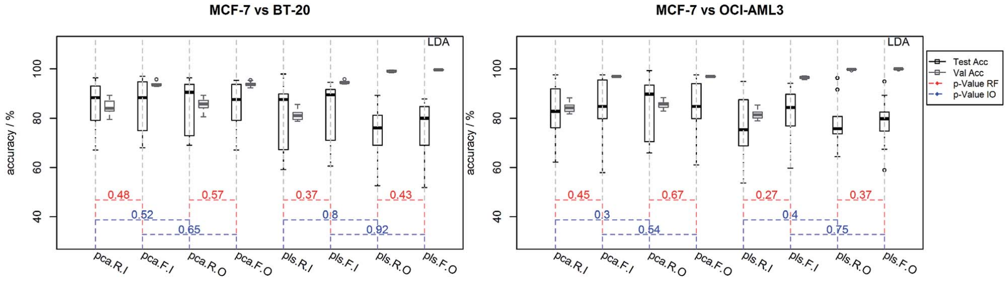


Fig. 4 Validation accuracies from the internal CV and the independent testing accuracies from the external CV. Hereby, the LDA was utilized for classification. The results of the SVM are shown in Fig. S4.† The applied data splitting methods, dimension reduction methods and the position for the internal CV are represented by the *x*-axis labels, where 'R' and 'F' represent the k-replicate CV and k-fold CV, respectively; while 'I' and 'O' denote the inside-CV and outside-CV, respectively. The validation and testing accuracies were consistent for the k-replicate inside-CV, which means that the model was evaluated reliably. In contrast, the validation accuracies were significantly higher than the testing accuracies for k-fold CV and outside-CV. This demonstrated an over-estimation of the model. However, the over-estimation of k-fold-CV and outside-CV was ignorable, if PCA was used for dimension reduction, demonstrated by the comparable validation and testing accuracies. To check the influence of the investigated mistakes of CV with respect to model optimization, we compared the testing accuracies for k-fold CV against k-replicate CV (RF) and inside-CV against outside-CV (IO). The comparison was done by applying a Wilcoxon-test. According to the *p*-values marked in the plot, no significant difference was observed. This means that the investigated two mistakes were less influential, if a CV was used for model parameter optimization.

k-fold CV doesn't reflect the real applications, where the dataset to be predicted is usually a different replicate compared with the training datasets. When it comes to the k-replicate CV, the validation dataset was a different replicate and not included in the training dataset. The procedure was like the real application scenario and the independence condition of CV was satisfied. Therefore, the dataset should be split at the highest hierarchical level to avoid the over-estimation of classification models.

It is noteworthy that the two data splitting mechanisms led to comparable results for PLS in the case of outside-CV, which is shown in Fig. 3. This results from the fact that PLS is a supervised method and the label information is already used during the model construction. For the outside-CV, both the training and validation datasets were used to create the dimension reduction model. Therefore, the information of the validation dataset was already utilized by the PLS model. Consequently, the validation dataset is not independent of the training dataset regardless of the used data splitting methods. The situation was different for PCA, where the data splitting mechanism made a large difference regardless of the position of PCA. In fact, the dataset splitting was more influential than the position of PCA. As previously explained, the outside-CV for PCA does not necessarily result in an over-estimation since no label information is used for PCA. Similarly, the eigenvectors of PCA do not necessarily vary with different data splitting methods. However, the influence of data splitting occurs during the following classifier training. With k-fold CV, the label information of the validation dataset was involved during the training of the classifier. This behaviour is similar to the PLS with outside-CV and definitely leads to an over-estimation compared to the k-replicate-CV.

Therefore, in the case of model evaluation, an improper dataset splitting always leads to severe mistakes and an unreliable model evaluation by CV. The second important point is the position of dimension reduction, especially for a supervised method like a PLS, which should be applied in the inside-CV.

Until now, we checked the validity of the internal CV in the case of model evaluation. The validation accuracy was used as a benchmark of the generalization performance of the model, while the testing accuracy was calculated to check the reliability of the benchmark. Another important aspect is the validity of the internal CV in the case of model parameter optimization. Here the internal CV was used to optimize the model parameters, the *n*PC or *n*LV in our investigation. The model was built with the optimal parameter based on the overall training dataset. The external CV, where the testing dataset was predicted, was used to verify the goodness of the optimization. In order to check the two mistakes of internal CV with respect to model optimization, we compared the testing accuracies that resulted from the different cases of the internal CV. The internal CV differed in data splitting methods, positions of the dimension reduction, and methods of dimension reduction. We compared the testing accuracies in the case of the 'PCA.R.I' and 'PCA.F.I', where the data splitting method was varied within the internal CV. Meanwhile the testing accuracies from the 'PCA.R.I' were compared to those from 'PCA.R.O' to check the influence of positions of PCA in the internal CV. The comparisons were done by a Wilcoxon-test. All the calculated *p*-values are higher than 0.05, as shown in Fig. 4, demonstrating no significant difference for all the comparisons. Therefore, the two common mistakes of CV are less influential for model optimization than for model evaluation. This is because for optimization the average validation accuracies for varying model parameters are compared. An over-estimation caused by mistakes in application of the CV increases all the validation accuracies over the possible values of the checked parameters. Therefore, the relative comparison between the parameters can still be meaningful, which finally leads to a satisfactory optimization.

Conclusions

In this contribution, we studied two common mistakes of cross-validating classification models. The first mistake is related to the data splitting methods, where we compared the k-replicate CV and k-fold CV. The second mistake relates to the position of the dimension reduction in terms of the CV loop. To do so, we compared the outside-CV and inside-CV for the dimension reduction methods including PCA and PLS. Two binary classification tasks were performed by LDA or SVM. A two-layer CV was utilized to study the CV for model evaluation and parameter optimization. The validation accuracies were calculated from the internal CV, while the testing accuracies were estimated by the external CV. As per these investigations, several important conclusions can be drawn and used as guidelines to avoid mistakes when applying a CV. These points are summarized in the following:

(1) If the internal CV is applied for model evaluation, the supervised dimension reduction techniques should always be included in the CV loop. Otherwise, the models are seriously over-estimated. As per unsupervised methods the over-estimation is less obvious, because the label information is not used to construct the dimension reduction projection. However, an inside-CV is still highly recommended.

(2) To correctly evaluate the classification models, the division of the dataset into training and validation datasets should always be carried out at the highest hierarchical level of the dataset such as biological/technical replicates.

(3) If the internal CV is used for optimizing model parameters, the above-mentioned two points are less influential. Nevertheless, caution is still recommended.

In conclusion, only if the above-mentioned points are carefully considered while applying a CV, the results from CV are reliable and robust. Only in this way a reliable model optimization and estimation can be carried out.

Acknowledgements

Financial support of the BMBF and EU *via* the project "Photo-Skin" (FKZ: 13N13243), the BMBF *via* the project "Intersept" (FKZ 13N13852), and the EU *via* the project "HemoSpec" (FP 7, CN 611682) is gratefully acknowledged. The scholarship for S. G. from the "China Scholarship Council" (CSC) is greatly appreciated.

References

1 M. Diem, *et al.*, Applications of infrared and Raman microspectroscopy of cells and tissue in medical diagnostics: present status and future promises, *J. Spectrosc.*, 2012, **27**(5–6), 463–496.

2 M. Diem, *et al.*, Molecular pathology *via* IR and Raman spectral imaging, *J. Biophotonics*, 2013, **6**(11–12), 855–886.

3 C. Krafft, *et al.*, Raman and coherent anti-Stokes Raman scattering microspectroscopy for biomedical applications, *J. Biomed. Opt.*, 2012, **17**(4), 0408011–04080115.

4 T. Bocklitz, *et al.*, Raman based molecular imaging and analytics: a magic bullet for biomedical applications!?, *Anal. Chem.*, 2016, **88**(1), 133–151.

5 A. C. S. Talari, *et al.*, Raman spectroscopy of biological tissues, *Appl. Spectrosc. Rev.*, 2015, **50**(1), 46–111.

6 C. Bielecki, *et al.*, Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells, *J. Biomed. Opt.*, 2012, **17**(7), 0760301–0760308.

7 E. Vargis, *et al.*, Detecting biochemical changes in the rodent cervix during pregnancy using Raman spectroscopy, *Ann. Biomed. Eng.*, 2012, **40**(8), 1814–1824.

8 W. Richardson, *et al.*, Ensemble multivariate analysis to improve identification of articular cartilage disease in noisy Raman spectra, *J. Biophotonics*, 2015, **8**(7), 555–566.

9 H. Abramczyk and B. Brozek-Pluska, Raman imaging in biochemical and biomedical applications. Diagnosis and treatment of breast cancer, *Chem. Rev.*, 2013, **113**(8), 5766–5781.

10 K. Kong, *et al.*, Raman spectroscopy for medical diagnostics—from *in vitro* biofluid assays to *in vivo* cancer detection, *Adv. Drug Delivery Rev.*, 2015, **89**, 121.

11 W. Wang, *et al.*, Real-time *in vivo* cancer diagnosis using Raman spectroscopy, *J. Biophotonics*, 2015, **8**(7), 527–545.

12 S. F. El-Mashtoly, *et al.*, Label-free imaging of drug distribution and metabolism in colon cancer cells by Raman microscopy, *Analyst*, 2014, **139**(5), 1155–1161.

13 U. Schmid, *et al.*, Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-Raman spectroscopy, *Chemom. Intell. Lab. Syst.*, 2009, **96**(2), 159–171.

14 M. Krause, *et al.*, The investigation of single bacteria by means of fluorescence staining and Raman spectroscopy, *J. Raman Spectrosc.*, 2007, **38**(4), 369–372.

15 M. Krause, *et al.*, Localizing and identifying living bacteria in *an abiotic* environment by a combination of Raman and fluorescence microscopy, *Anal. Chem.*, 2008, **80**(22), 8568–8575.

16 M. Jermyn, *et al.*, Intraoperative brain cancer detection with Raman spectroscopy in humans, *Sci. Transl. Med.*, 2015, 7(274), 274ra19.

17 T. Bocklitz, *et al.*, A comprehensive study of classification methods for medical diagnosis, *J. Raman Spectrosc.*, 2009, **40**(12), 1759–1765.

18 M. Browne, Cross-validation methods, *J. Math. Psychol.*, 2000, **44**(1), 108–132.

19 T. Hastie, R. Tibshirani and J. Friedman, *Model Assessment and Selection, the Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer, 2nd edn, 2008.

20 R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in *Ijcai*, 1995.

21 S. Arlot and A. Celisse, A survey of cross-validation procedures for model selection, *Statistics Surveys*, 2010, **4**, 40–79.

This journal is © The Royal Society of Chemistry 2017

22 J. A. Westerhuis, *et al.*, Assessment of PLSDA cross validation, *Metabolomics*, 2008, **4**(1), 81–89.

23 M. Häfner, *et al.*, Evaluation of cross-validation protocols for the classification of endoscopic images of colonic polyps, in *Computer-based Medical Systems (CBMS), 2012 25th International Symposium on*, IEEE, 2012.

24 P. Filzmoser, B. Liebmann and K. Varmuza, Repeated double cross validation, *J. Chemom.*, 2009, **23**(4), 160–171.

25 M. Defernez and E. K. Kemsley, The use and misuse of chemometrics for treating classification problems, *TrAC, Trends Anal. Chem.*, 1997, **16**(4), 216–221.

26 F. R. Burden, R. G. Brereton and P. T. Walsh, Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy, *Analyst*, 1997, **122**(10), 1015–1022.

27 C. Beleites, *et al.*, Sample size planning for classification models, *Anal. Chim. Acta*, 2013, **760**, 25–33.

28 developers, s., {signal}: Signal processing. 2014.

29 K. H. Liland and B.-H. Mevik, *Baseline: Baseline Correction of Spectra*, 2015.

30 W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Springer, New York, 4th edn, 2002.

31 M. David, *et al.*, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*, 2015.

32 R. L. McCreery, *Raman Spectroscopy for Chemical Analysis*, John Wiley & Sons, 2000, vol. 157.

33 E. V. Thomas, Incorporating auxiliary predictor variation in principal component regression models, *J. Chemom.*, 1995, **9**(6), 471–481.

This journal is © The Royal Society of Chemistry 2017

Journal Name

ARTICLE

Common Mistakes in Cross-Validating Classification Models

Shuxia Guo,[a, b, †] Thomas Bocklitz,[a, b, †] Ute Neugebauer,[a, b, c] and Jürgen Popp [a, b,d]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

In this contribution we investigated the common mistakes of cross-validation (CV) for the development of chemometric models for Raman based biological applications. We focused on two common mistakes: the first mistake occurs when splitting the dataset into training and validation data sets improperly; and the second mistake is regarding the wrong position of a dimension reduction procedure with respect to the CV loop. For the first mistake, we split the dataset either randomly or each technical replicate was used as one fold of the CV and compared the results. To check the second mistake, we employed two dimension reduction methods including principal component analysis (PCA) and partial least squares regression (PLS). These dimension reduction models were constructed either once for the whole training data outside the CV loop or rebuilt inside the CV loop for each iteration. We based our study on a benchmark dataset of Raman spectra of three cell types (MCF-7, BT-20, and OCI-AML3), which included nine technical replicates respectively. Two binary classification models were constructed with a two-layer CV. For the external CV, each replicate was used once as the independent testing data set. The other replicates were used for the internal CV, where different methods of data splitting and different positions of the dimension reduction were studied.

The conclusions include two points. The first point is related to the reliability of the model evaluation by the internal CV, illustrated by the differences between the testing accuracies from the external CV and the validation accuracies from the internal CV. It was demonstrated that the dataset should be split at the highest hierarchical level, which means the biological/technical replicate in this manuscript. Meanwhile, the dimension reduction should be redone each iteration of the internal CV loop. The second aspect relates to the optimization performance of the internal CV, benchmarked by the prediction accuracy of the optimized model on the testing data set. Comparable results were observed for different methods of data splitting and positions of dimension reduction in the internal CV. That means if the internal CV is used for optimizing the model parameters, the two mistakes are less influential in contrast to the model evaluation.

---

[a] *Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany.*
[b] *Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany.*
[c] *Center for Sepsis Control and Care, Jena University Hospital, Germany.*
[d] *InfectoGnostics Research Campus Jena, Center for Applied Research, Jena, Germany.*

† *These authors share main authorship due to equal contributions.*

Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

Journal Name

ARTICLE

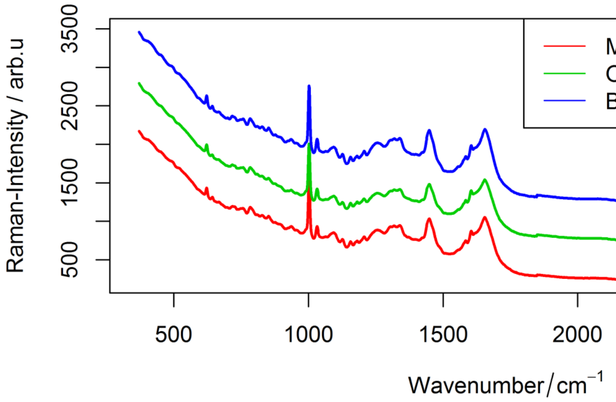


Fig. S1. Mean Raman spectra of the investigated three cell types.

```
////inside-CV
for i=1:N    //external CV
  $ts_i$=data without ith replicate
  $vs_i$=data of ith replicate
  fold=split($ts_i$, replicate, nFold=N-1)  //split randomly or as replicate
  for nDim=3:50
    for j=1:(N-1) //internal CV
      $ts_{ij}$=$ts_i$[fold[-j],]   //jth training set for ith iteration of external CV
      $vs_{ij}$=$ts_i$[fold[j],]   //jth validation set for ith iteration of external CV
      pca=PCA($ts_{ij}$, nDim)
      classifier=LDA(predict(pca, $ts_{ij}$))
      $acc_{nj}$=accuracy(predict(classifier, predict(pca, $vs_{ij}$)))
    End
  End
  optDim=nDim with the maximum in rowMeans(acc)
  pca=PCA($ts_i$, optDim)
  $acc_i$=accuracy(predict(classifier, predict(pca, $vs_i$)))
End
```

```
////outside-CV
for i=1:N    //external CV
  $ts_i$=data without ith replicate
  $vs_i$=data of ith replicate
  fold=split($ts_i$, replicate, nFold=N-1)  //split randomly or as replicate
  for nDim=3:50
    scores=predict(PCA($ts_i$, nDim),$ts_i$)
    for j=1:(N-1) //internal validation
      $tIndex_{ij}$=fold[-j]   //jth training set for ith iteration of external CV
      $vIndex_{ij}$=fold[j]   //jth validation set for ith iteration of external CV
      classifier=LDA(scores[$tIndex_{ij}$,])
      $acc_{nj}$=accuracy(predict(classifier, scores[$tIndex_{ij}$,]))
    End
  End
  optDim=nDim with the maximum in rowMeans(acc)
  pca=PCA($ts_i$, optDim)
  $acc_i$=accuracy(predict(classifier, predict(pca, $vs_i$)))
End
```

Fig. S2. . Pseudo code of the applied two-layer CV. Models with different component numbers *nPC* (*nLV*) were built and validated with an internal CV. Each replicate was taken out once and predicted within the external CV. For each iteration of the external CV, the model was built based on the overall training set with the *nPC* (*nLV*) featuring the highest averaged validation accuracy. (1) Within the Inside-CV, a dimension reduction method (PCA/PLS) was redone each iteration of the internal CV loop. Thus the PCA or PLS was executed after removing the validation set. The scores of the validation sets were predicted and then classified by the classification model (LDA or SVM). (2) For the Outside-CV, the dimension reduction method was carried out once for all data outside the internal CV loop. Therefore the validation set was involved in constructing the PCA/PLS model. Afterwards, the scores were split into training and validation sets for internal CV.

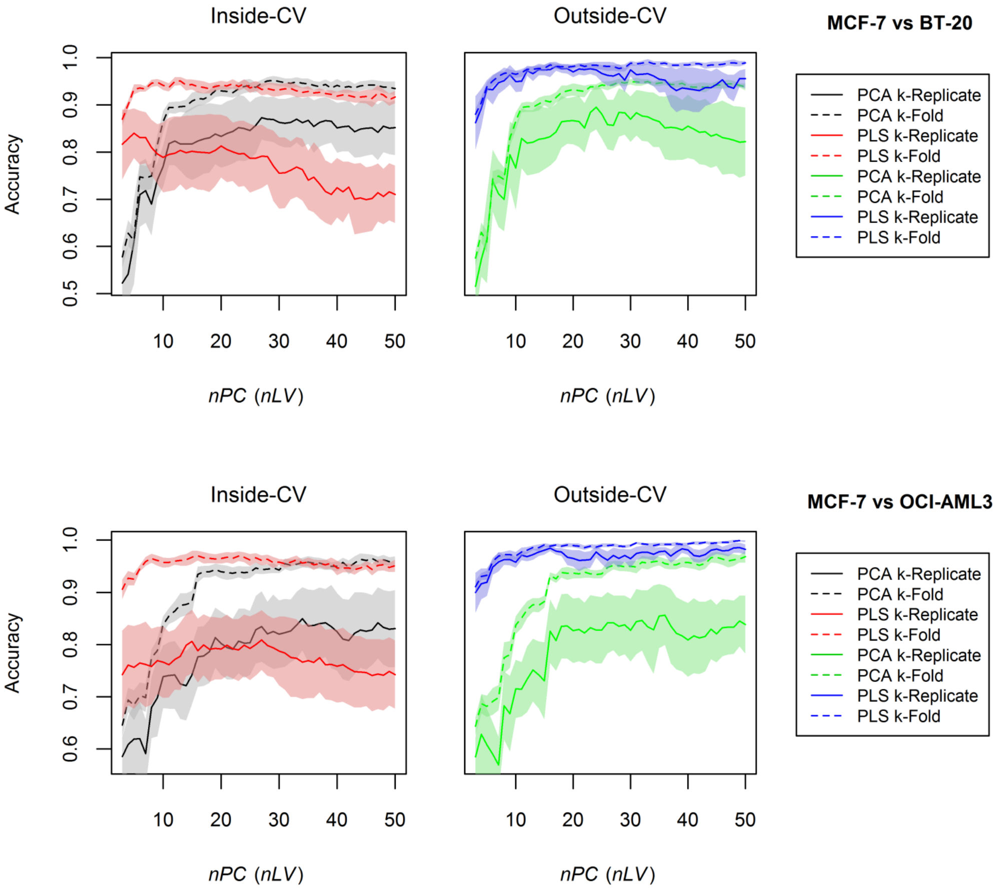


Fig. S3. Validation accuracies of the internal CV within the first iteration of the external CV, for two binary classification tasks (MCF-7 vs BT-20 (m vs b) and MCF-7 vs OCI-AML3 (m vs o)) based on SVM. In both tasks, the outside-CV yields higher accuracies. This is more obvious if supervised dimension reduction methods, such as PLS, are applied. The validation accuracies are always higher if a k-fold CV is used compared with a k-replicate CV. The over-estimation of the k-fold CV is due to the violated independence criteria between the training and validation sets. This effect of over-estimation is more enhanced for supervised dimension reduction.

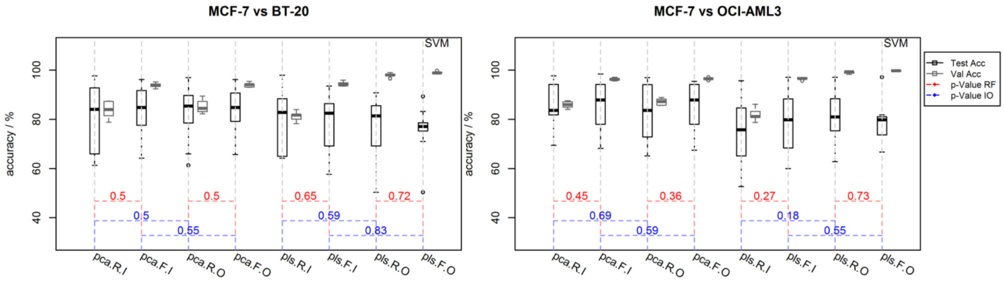


Fig. S4. Validation accuracies resulted from the internal CV and the independent testing accuracies from the external CV. Hereby the SVM was utilized for classification. The applied data splitting methods, dimension reduction methods and the position for the internal CV are referred to the x-axis labels, where ‘R’ and ‘F’ represent the k-replicate CV and k-fold CV, respectively; while ‘I’ and ‘O’ denote the inside-CV and outside-CV, respectively. The validation and testing accuracies were consistent for the k-replicate inside-CV, which means the model was evaluated reliably. On the contrary, the validation accuracies were significantly higher than the testing accuracies for k-fold CV and outside-CV. This demonstrated an over-estimation of the model. However, the over-estimation of k-fold-CV and outside-CV was ignorable if PCA was used for dimension reduction, demonstrated by the comparable validation and testing accuracies. In addition, in order to check the influence of the investigated two mistakes of CV with respective of model optimization, we compared the testing accuracies for k-fold CV against k-replicate CV (RF), and inside-CV against outside-CV (IO). The comparison was done by Wilcoxon-test. According to the p-values marked in the plot, no significant difference was observed. That means the investigated two mistakes were less influential if CV was used for model parameter optimization.

7.4 Towards an Improvement of Model Transferability for Raman Spectroscopy in Biological Applications (A4)

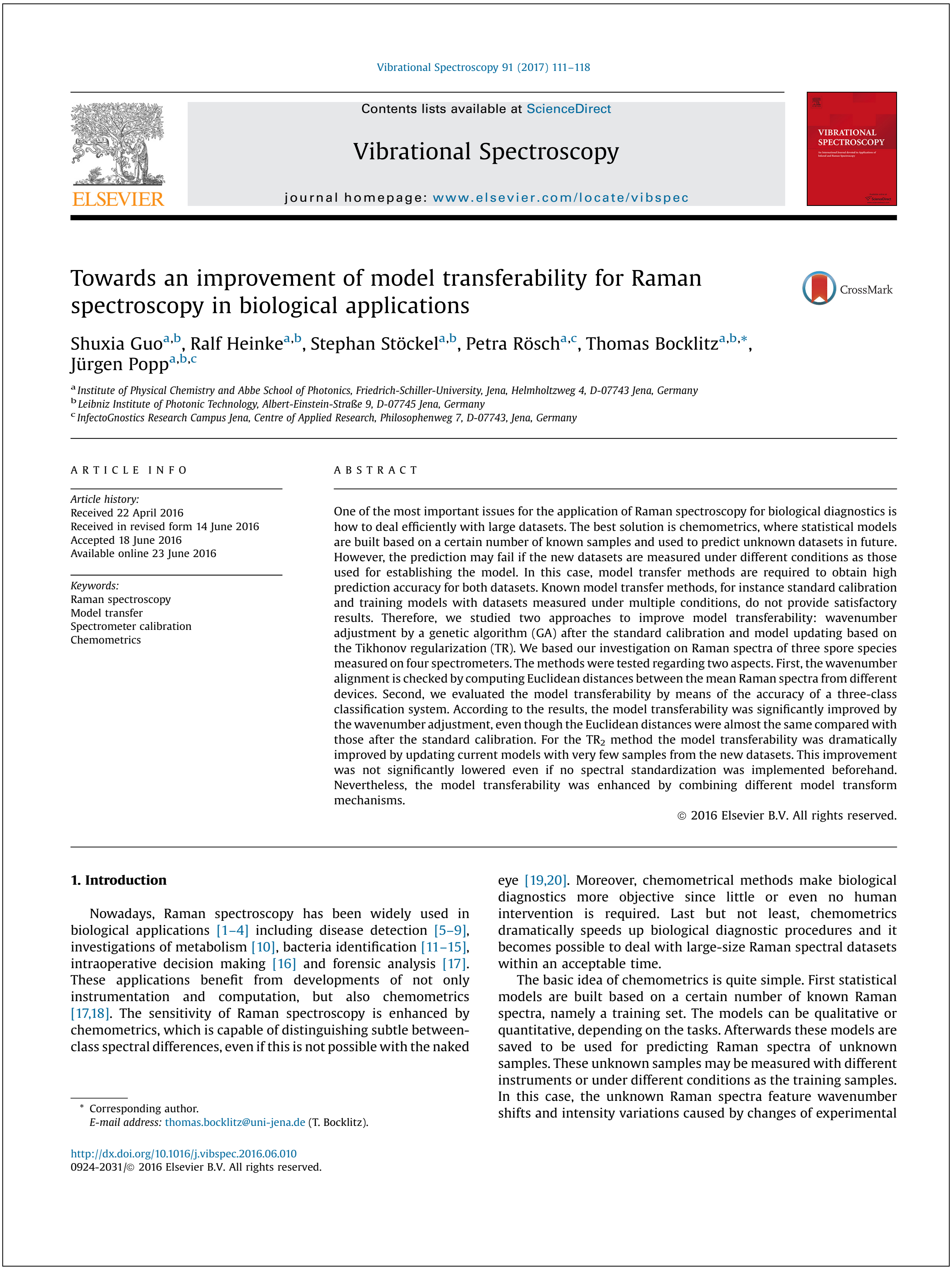
S. Guo, R. Heinke, S. Stöckel, P. Rösch, T. Bocklitz, and J. Popp, *Vibrational Spectroscopy*, 2017, 91, 111-118.

Der Nachdruck der folgenden Publikation erscheint mit freundlicher Genehmigung von ELSEVIER. Reprinted with kind permission from ELSEVIER.

Erklärungen zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation

Towards an improvement of model transferability for Raman spectroscopy in biological applications, S. Guo[1], R. Heinke[2], S. Stöckel[3], P. Rösch[4], T. Bocklitz[5], and J. Popp[6], *Vibrational Spectroscopy*, 2017, 91, 111-118.

Beteiligt an (*Zutreffendes ankreuzen*)						
	1	2	3	4	5	6
Konzeption des Forschungsansatzes	x				x	x
Planung der Untersuchungen	x	x	x	x	x	x
Datenerhebung		x	x	x		
Datenanalyse und -interpretation	x				x	
Schreiben des Manuskripts	x		x	x	x	x
Vorschlag Anrechnung Publikationsäquivalente	1.0					



Vibrational Spectroscopy 91 (2017) 111–118

Contents lists available at ScienceDirect

Vibrational Spectroscopy

journal homepage: www.elsevier.com/locate/vibspec

Towards an improvement of model transferability for Raman spectroscopy in biological applications

Shuxia Guo[a,b], Ralf Heinke[a,b], Stephan Stöckel[a,b], Petra Rösch[a,c], Thomas Bocklitz[a,b,*], Jürgen Popp[a,b,c]

[a] *Institute of Physical Chemistry and Abbe School of Photonics, Friedrich-Schiller-University, Jena, Helmholtzweg 4, D-07743 Jena, Germany*
[b] *Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07745 Jena, Germany*
[c] *InfectoGnostics Research Campus Jena, Centre of Applied Research, Philosophenweg 7, D-07743, Jena, Germany*

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form 14 June 2016
Accepted 18 June 2016
Available online 23 June 2016

Keywords:
Raman spectroscopy
Model transfer
Spectrometer calibration
Chemometrics

ABSTRACT

One of the most important issues for the application of Raman spectroscopy for biological diagnostics is how to deal efficiently with large datasets. The best solution is chemometrics, where statistical models are built based on a certain number of known samples and used to predict unknown datasets in future. However, the prediction may fail if the new datasets are measured under different conditions as those used for establishing the model. In this case, model transfer methods are required to obtain high prediction accuracy for both datasets. Known model transfer methods, for instance standard calibration and training models with datasets measured under multiple conditions, do not provide satisfactory results. Therefore, we studied two approaches to improve model transferability: wavenumber adjustment by a genetic algorithm (GA) after the standard calibration and model updating based on the Tikhonov regularization (TR). We based our investigation on Raman spectra of three spore species measured on four spectrometers. The methods were tested regarding two aspects. First, the wavenumber alignment is checked by computing Euclidean distances between the mean Raman spectra from different devices. Second, we evaluated the model transferability by means of the accuracy of a three-class classification system. According to the results, the model transferability was significantly improved by the wavenumber adjustment, even though the Euclidean distances were almost the same compared with those after the standard calibration. For the TR_2 method the model transferability was dramatically improved by updating current models with very few samples from the new datasets. This improvement was not significantly lowered even if no spectral standardization was implemented beforehand. Nevertheless, the model transferability was enhanced by combining different model transform mechanisms.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

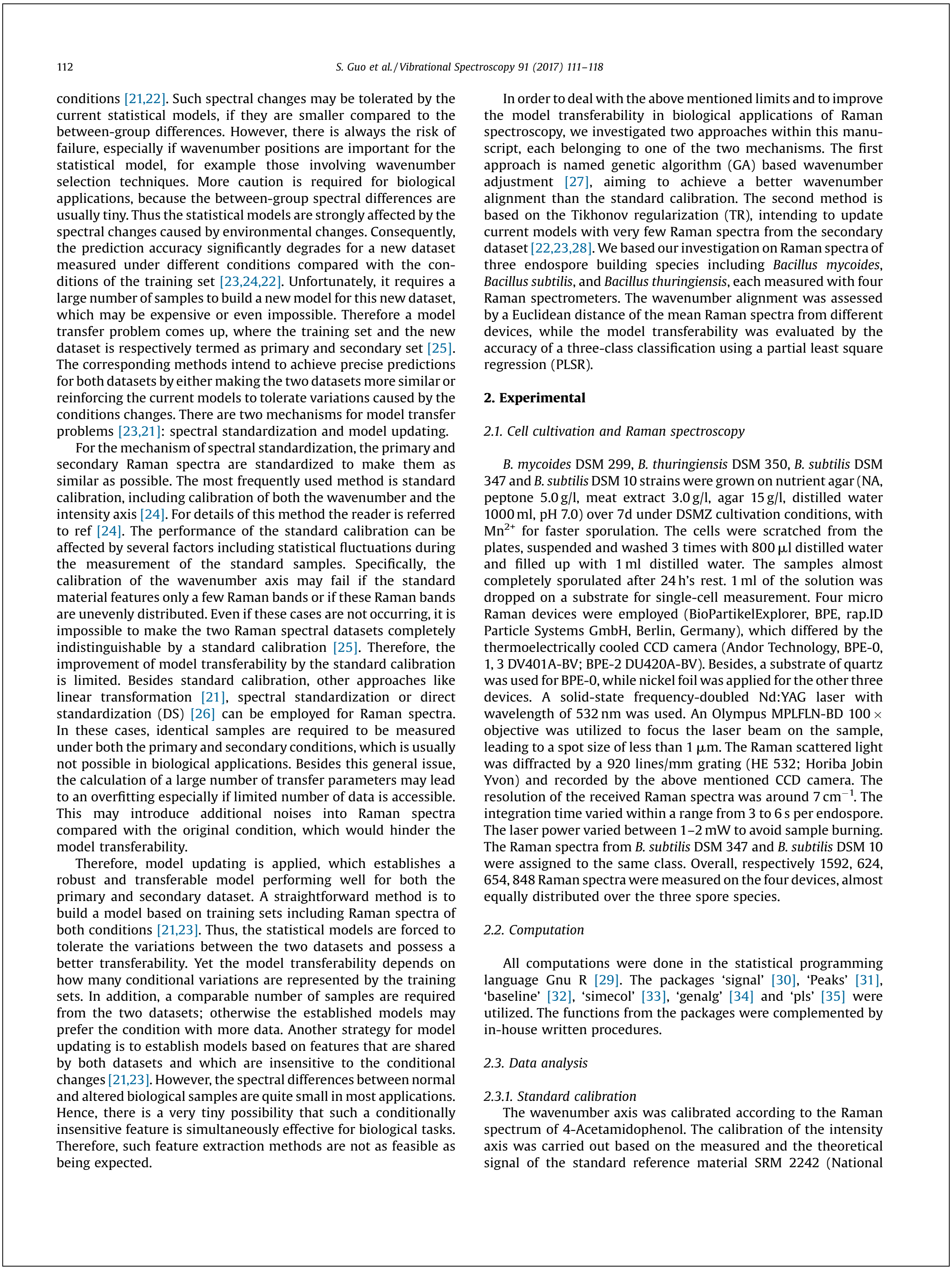
Nowadays, Raman spectroscopy has been widely used in biological applications [1–4] including disease detection [5–9], investigations of metabolism [10], bacteria identification [11–15], intraoperative decision making [16] and forensic analysis [17]. These applications benefit from developments of not only instrumentation and computation, but also chemometrics [17,18]. The sensitivity of Raman spectroscopy is enhanced by chemometrics, which is capable of distinguishing subtle between-class spectral differences, even if this is not possible with the naked eye [19,20]. Moreover, chemometrical methods make biological diagnostics more objective since little or even no human intervention is required. Last but not least, chemometrics dramatically speeds up biological diagnostic procedures and it becomes possible to deal with large-size Raman spectral datasets within an acceptable time.

The basic idea of chemometrics is quite simple. First statistical models are built based on a certain number of known Raman spectra, namely a training set. The models can be qualitative or quantitative, depending on the tasks. Afterwards these models are saved to be used for predicting Raman spectra of unknown samples. These unknown samples may be measured with different instruments or under different conditions as the training samples. In this case, the unknown Raman spectra feature wavenumber shifts and intensity variations caused by changes of experimental

---

[*] Corresponding author.
E-mail address: thomas.bocklitz@uni-jena.de (T. Bocklitz).

http://dx.doi.org/10.1016/j.vibspec.2016.06.010
0924-2031/© 2016 Elsevier B.V. All rights reserved.



conditions [21,22]. Such spectral changes may be tolerated by the current statistical models, if they are smaller compared to the between-group differences. However, there is always the risk of failure, especially if wavenumber positions are important for the statistical model, for example those involving wavenumber selection techniques. More caution is required for biological applications, because the between-group spectral differences are usually tiny. Thus the statistical models are strongly affected by the spectral changes caused by environmental changes. Consequently, the prediction accuracy significantly degrades for a new dataset measured under different conditions compared with the conditions of the training set [23,24,22]. Unfortunately, it requires a large number of samples to build a new model for this new dataset, which may be expensive or even impossible. Therefore a model transfer problem comes up, where the training set and the new dataset is respectively termed as primary and secondary set [25]. The corresponding methods intend to achieve precise predictions for both datasets by either making the two datasets more similar or reinforcing the current models to tolerate variations caused by the conditions changes. There are two mechanisms for model transfer problems [23,21]: spectral standardization and model updating.

For the mechanism of spectral standardization, the primary and secondary Raman spectra are standardized to make them as similar as possible. The most frequently used method is standard calibration, including calibration of both the wavenumber and the intensity axis [24]. For details of this method the reader is referred to ref [24]. The performance of the standard calibration can be affected by several factors including statistical fluctuations during the measurement of the standard samples. Specifically, the calibration of the wavenumber axis may fail if the standard material features only a few Raman bands or if these Raman bands are unevenly distributed. Even if these cases are not occurring, it is impossible to make the two Raman spectral datasets completely indistinguishable by a standard calibration [25]. Therefore, the improvement of model transferability by the standard calibration is limited. Besides standard calibration, other approaches like linear transformation [21], spectral standardization or direct standardization (DS) [26] can be employed for Raman spectra. In these cases, identical samples are required to be measured under both the primary and secondary conditions, which is usually not possible in biological applications. Besides this general issue, the calculation of a large number of transfer parameters may lead to an overfitting especially if limited number of data is accessible. This may introduce additional noises into Raman spectra compared with the original condition, which would hinder the model transferability.

Therefore, model updating is applied, which establishes a robust and transferable model performing well for both the primary and secondary dataset. A straightforward method is to build a model based on training sets including Raman spectra of both conditions [21,23]. Thus, the statistical models are forced to tolerate the variations between the two datasets and possess a better transferability. Yet the model transferability depends on how many conditional variations are represented by the training sets. In addition, a comparable number of samples are required from the two datasets; otherwise the established models may prefer the condition with more data. Another strategy for model updating is to establish models based on features that are shared by both datasets and which are insensitive to the conditional changes [21,23]. However, the spectral differences between normal and altered biological samples are quite small in most applications. Hence, there is a very tiny possibility that such a conditionally insensitive feature is simultaneously effective for biological tasks. Therefore, such feature extraction methods are not as feasible as being expected.

In order to deal with the above mentioned limits and to improve the model transferability in biological applications of Raman spectroscopy, we investigated two approaches within this manuscript, each belonging to one of the two mechanisms. The first approach is named genetic algorithm (GA) based wavenumber adjustment [27], aiming to achieve a better wavenumber alignment than the standard calibration. The second method is based on the Tikhonov regularization (TR), intending to update current models with very few Raman spectra from the secondary dataset [22,23,28]. We based our investigation on Raman spectra of three endospore building species including *Bacillus mycoides*, *Bacillus subtilis*, and *Bacillus thuringiensis*, each measured with four Raman spectrometers. The wavenumber alignment was assessed by a Euclidean distance of the mean Raman spectra from different devices, while the model transferability was evaluated by the accuracy of a three-class classification using a partial least square regression (PLSR).

2. Experimental

2.1. Cell cultivation and Raman spectroscopy

B. mycoides DSM 299, *B. thuringiensis* DSM 350, *B. subtilis* DSM 347 and *B. subtilis* DSM 10 strains were grown on nutrient agar (NA, peptone 5.0 g/l, meat extract 3.0 g/l, agar 15 g/l, distilled water 1000 ml, pH 7.0) over 7d under DSMZ cultivation conditions, with Mn^{2+} for faster sporulation. The cells were scratched from the plates, suspended and washed 3 times with 800 µl distilled water and filled up with 1 ml distilled water. The samples almost completely sporulated after 24 h's rest. 1 ml of the solution was dropped on a substrate for single-cell measurement. Four micro Raman devices were employed (BioPartikelExplorer, BPE, rap.ID Particle Systems GmbH, Berlin, Germany), which differed by the thermoelectrically cooled CCD camera (Andor Technology, BPE-0, 1, 3 DV401A-BV; BPE-2 DU420A-BV). Besides, a substrate of quartz was used for BPE-0, while nickel foil was applied for the other three devices. A solid-state frequency-doubled Nd:YAG laser with wavelength of 532 nm was used. An Olympus MPLFLN-BD 100 × objective was utilized to focus the laser beam on the sample, leading to a spot size of less than 1 µm. The Raman scattered light was diffracted by a 920 lines/mm grating (HE 532; Horiba Jobin Yvon) and recorded by the above mentioned CCD camera. The resolution of the received Raman spectra was around 7 cm^{-1}. The integration time varied within a range from 3 to 6 s per endospore. The laser power varied between 1–2 mW to avoid sample burning. The Raman spectra from *B. subtilis* DSM 347 and *B. subtilis* DSM 10 were assigned to the same class. Overall, respectively 1592, 624, 654, 848 Raman spectra were measured on the four devices, almost equally distributed over the three spore species.

2.2. Computation

All computations were done in the statistical programming language Gnu R [29]. The packages 'signal' [30], 'Peaks' [31], 'baseline' [32], 'simecol' [33], 'genalg' [34] and 'pls' [35] were utilized. The functions from the packages were complemented by in-house written procedures.

2.3. Data analysis

2.3.1. Standard calibration

The wavenumber axis was calibrated according to the Raman spectrum of 4-Acetamidophenol. The calibration of the intensity axis was carried out based on the measured and the theoretical signal of the standard reference material SRM 2242 (National

Institute of Standards & Technology, Gaithersburg, MD 20899, USA).

2.3.2. Pre-processing

Raman spectra were interpolated to an equidistant wavenumber grid of 1 cm^{-1} and smoothed by a 2-order Savitzky-Golay filtering with a window width of 11. Baseline correction was performed by the asymmetric least squares (ALS) method in R package ‘baseline’ (lambda = 7, p = 0.01). Vector normalization was carried out to remove the interference of variation of the integration time.

2.3.3. Classification

A three-class classification was performed by partial least square regression (PLSR) based on the first fifteen components. The averaged accuracy of a 5-fold cross-validation was employed as an evaluation of the performance of models.

2.3.4. Genetic algorithm

As it was mentioned above, the standard calibration cannot eliminate completely the spectral differences between the primary and secondary datasets. This leads to limited model transferability. To improve the result, we attempted to obtain a better wavenumber alignment, which is described in this section. To be differentiated from the standard wavenumber calibration, this procedure is termed as wavenumber adjustment in the manuscript.

From a mathematic point of view, the task is similar with the standard wavenumber calibration. The Raman spectra of primary and secondary datasets are aligned to a standard spectrum. This standard Raman spectrum could be the mean Raman spectrum of either the primary or secondary dataset, or both datasets. In principle, the wavenumber alignment can be realized by an identical process as the standard wavenumber calibration. That is to say, the adjustment over the whole wavenumber axis is obtained according to certain positions such as the positions of Raman bands. However, in this way the final result is optimal only for these certain positions instead of the overall wavenumber axis. This may lead to a sub-optimal alignment. To deal with this issue, we developed a genetic algorithm, aiming to search for wavenumber adjustments resulting in the highest similarity between the Raman spectrum under-investigation and the standard spectrum. Fig. 1 shows the workflow. All Raman spectra should be pre-processed at first, but it does not matter whether or not the standard calibration is implemented beforehand.

To begin with, the Raman spectra of the primary and secondary datasets are averaged to generate the standard spectrum, termed as reference spectrum. Simultaneously, the respective mean Raman spectrum of the two datasets is calculated and marked as target spectrum. The wavenumber alignment is executed between the target spectrum and the standard spectrum. To avoid overfitting, the Raman peak positions are located on the reference spectrum, marked as control points. The adjustment of wavenumber at each control point from the target to the reference spectrum is represented by a gene. Thus the length of each chromosome is the same as the number of control points. The numerical range of the genes is from −10 to 10 cm^{-1}. The first generation is created randomly, with a population size of 20. With

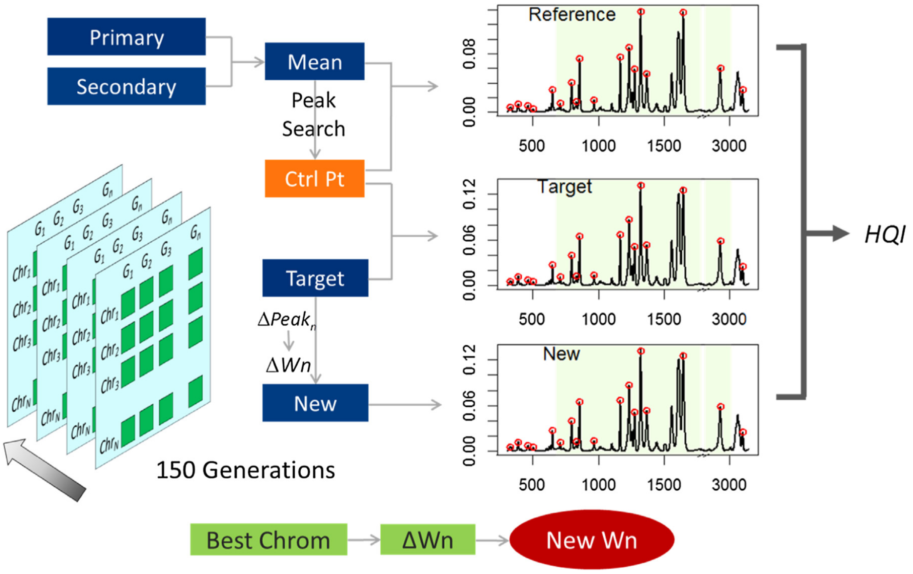


Fig. 1. Workflow of a genetic algorithm based wavenumber adjustment: To begin with, the primary and secondary datasets were averaged to generate a reference spectrum. Peaks were localized in this reference to receive the control points. Mean Raman spectra were computed separately for primary and secondary datasets and used as target spectra. Each gene represents an estimation of wavenumber adjustment at one control point, ranging from −10 to 10. The first population with 20 chromosomes is created randomly. The values of each chromosome are fitted to the whole wavenumber axis. Accordingly, the target Raman spectrum can be corrected similarly as standard wavenumber calibration, generating a new Raman spectrum. The hit-quality-index (HQI) between this new spectrum and the reference spectrum was calculated, where only the shaded spectral region was used. The algorithm evolved for 150 generations to increase the HQI value (Eq. (1)). The best chromosome in the last generation was used for wavenumber adjustment of all Raman spectra measured under the same condition as the target.

each chromosome, we receive a new wavenumber axis by fitting the values of the genes, i.e., the wavenumber adjustments of the control points, onto the whole wavenumber axis similarly as the standard wavenumber calibration. Afterwards, the hit-quality-index (HQI) [36] between the reference and the new Raman spectrum is calculated according to Eq. (1). This HQI is used as the evaluation function, which should be maximized during the algorithm evolvement. The best chromosome in the last generation is applied for the wavenumber adjustment of all Raman spectra measured under the same condition as the target spectrum.

$$HQI = \frac{(\boldsymbol{S}\cdot\boldsymbol{S}_{\text{ref}})^2}{\boldsymbol{S}^2\cdot\boldsymbol{S}_{\text{ref}}^2} \qquad (1)$$

In our experiment, the wavenumber alignment was employed after the standard calibration and pre-processing. The calculations of the new wavenumber axis were carried out on the Raman spectra of 4-Acetamidophenol measured by the four devices. The mean Raman spectrum of all devices was used as the reference spectrum. The mean spectrum of each device was employed as the target spectrum. The flat and noisy wavenumber range from 1785 to 2815 cm^{-1} was excluded during the computation of HQI. Thus only the shaded regions shown in Fig. 1 were used. Finally, the Raman spectra of the spore species, after standard calibration and pre-processing, were interpolated according to the corresponding new wavenumber axis. The genetic algorithm stopped after 150 generations. During the evolution, the best 20% parent individuals are kept in the next generation. The other 80% of child individuals are obtained by cross-over between any two of the parent individuals, with a mutation rate of 0.01.

It is noteworthy that this wavenumber alignment can also be utilized in cases where the standard calibration is not accessible, by performing the calculations on the pre-processed Raman spectra of real samples.

2.3.5. Tikhonov regularization

One of the drawbacks of spectral standardization methods is the risk of introducing artefacts into Raman spectra. This can be avoided with the other model transfer mechanism, which aims to update current statistical models with secondary datasets, as shown by Eq. (2). Here $\boldsymbol{X}$ is a $m \times n$ Raman spectral matrix with spectra in rows, representing the training set from the primary dataset. $\boldsymbol{y}$ is the corresponding output matrix of $\boldsymbol{X}$, with the size of $m \times k$ for a k-class classification. $\boldsymbol{y}_{*j}$ equals one if the Raman spectra belong to class j and otherwise $\boldsymbol{y}_{*j}$ is zero. $\boldsymbol{L}$ and $\boldsymbol{y}^*$ is respectively a matrix composed of Raman spectra from the secondary dataset and the output matrix.

$$\begin{pmatrix} y \\ \mathbf{y}^* \end{pmatrix} = \begin{pmatrix} X \\ L \end{pmatrix}\mathbf{b} \qquad (2)$$

Due to measurement expenses, there are usually more data used to construct $\boldsymbol{X}$ compared with the construction of the $\boldsymbol{L}$ matrix. This leads to a bad performance of the updated models when predicting the secondary data. Therefore, Eq. (2) is further developed into Eq. (3) by multiplying a numerical parameter λ to the $\boldsymbol{L}$ matrix. In this case, the secondary dataset possesses a larger weight than the primary dataset, which to some extent corrects for the lower sample size in the $\boldsymbol{L}$ matrix. Eq. (3) is a basic form of Tikhonov regularization (TR), which is referred as TR_1.

$$\begin{pmatrix} y \\ \lambda\mathbf{y}^* \end{pmatrix} = \begin{pmatrix} X \\ \lambda L \end{pmatrix}\mathbf{b} \qquad (3)$$

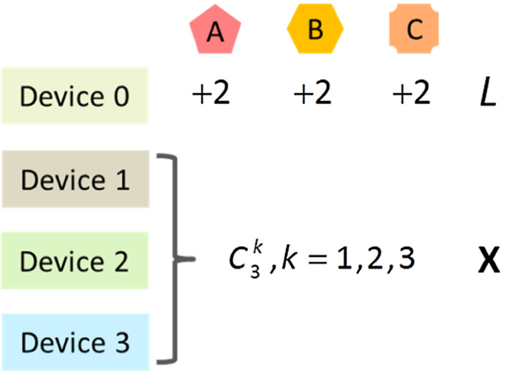


Fig. 2. Framework for forming matrices within TR: $\boldsymbol{L}$ was composed of six Raman spectra from the device to be predicted, each two belonging to one species. $\boldsymbol{X}$ was composed of Raman spectra measured by any one of the seven combinations of the other three devices. Therefore, Raman spectra from each device were predicted by seven models corresponding to seven $\boldsymbol{X}$ matrices.

Despite of the improvement, there are still drawbacks for TR_1. The matrix $\boldsymbol{L}$ is non-diagonal and may be collinear with $\boldsymbol{X}$, which would enhance the singularity of the multivariate calibration in Eq. (3) and eventually decrease the stability of the model [28]. To tackle this issue, an additional term η is introduced, as shown in Eq. (4). Here $\boldsymbol{I}$ is an identity matrix used to ensure the non-singularity of the calculation. This is another form of TR, termed as TR_2.

$$\begin{pmatrix} y \\ 0 \\ \lambda\mathbf{y}^* \end{pmatrix} = \begin{pmatrix} X \\ \eta I \\ \lambda L \end{pmatrix}\mathbf{b} \qquad (4)$$

In terms of these two TR forms, the first problem is how to construct the related matrices. Fig. 2 shows the basic procedure in our experiment. Assuming Raman spectra measured with device 0 are the secondary data to be predicted. Then six Raman spectra were selected randomly from this dataset to construct $\boldsymbol{L}$, each two belonging to one species. The training set, or matrix $\boldsymbol{X}$, was composed of Raman spectra from any possible combinations of the other three devices. Therefore, there were totally $C_3^1 + C_3^2 + C_3^3 = 7$ possible training sets, which resulted in seven models to predict Raman spectra from device 0. The prediction of datasets from the other three devices was performed similarly. The second problem of TR methods is the optimization of the parameters, which is described in the 'Results and discussion' section of the manuscript.

3. Results and discussion

Fig. 3 shows the mean Raman spectra of *B. mycoides* measured with the four devices, before and after the standard calibration. To make it clearer, we displayed the marked region in a zoomed version on the right side of Fig. 3. Apparently, the spectral differences among the four mean Raman spectra were decreased by the standard calibration. However, we can still observe obvious wavenumber and intensity variations. This proves the limit of the standard calibration to remove spectral variations originating from the measurement condition. The situation was similar for the mean Raman spectra of the other two species, which are not shown herein. A large difference was observed between BPE0 and the other three devices because of the different substrates utilized.

3.1. Euclidean distance

Fig. 3 provides an intuitive assessment of the spectral alignment. To quantify this, we computed the Euclidean distance as a quantitative evaluation of the spectral standardization.

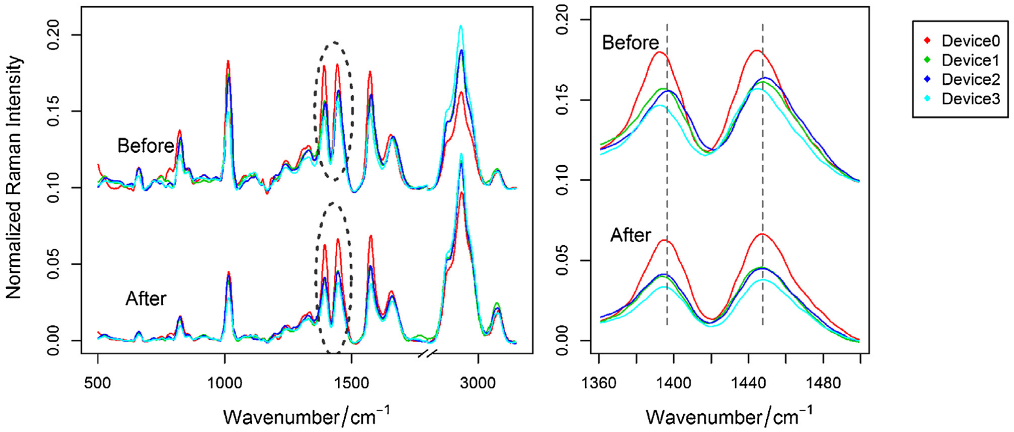


Fig. 3. Mean Raman spectra of *B. mycoides* measured with four devices, before and after the standard calibration: The marked regions are zoomed in the right plot. The spectral differences between devices are decreased after calibration, but far from being completely removed, which is clearer in the zoomed plot. This proved that the standard calibration cannot completely remove spectral variations caused by conditional changes. The large difference was observed between BPE0 and the other three devices because of the different substrates used.

Specifically, the mean Raman spectrum was computed for each species and each device. Euclidean distances were calculated between the mean Raman spectra belonging to the same species but measured on different devices. This calculation was done between each two out of the four devices, yielding six distance values for each species. We calculated these distances of Raman spectra before spectral standardization, after the standard calibration, and after the standard calibration plus the wavenumber adjustment. The Raman spectra were always preprocessed in the same manner.

The results are shown in Fig. 4, where the species are encoded with different colors. Generally, the results are quite similar among different species. The Raman spectra without any spectral standardization display a serious spectral discrepancy, which is reflected not only by high mean values of the Euclidean distances, but also by large variations. Such large variations indicate considerable between-device spectral differences. The spectral discrepancy was decreased after the standard calibration. However, the following wavenumber adjustment did not make further improvement. This is because the most part of spectral variations had been removed by the standard calibration, while the genetic algorithm only adjusted the wavenumber axis in a slight extend.

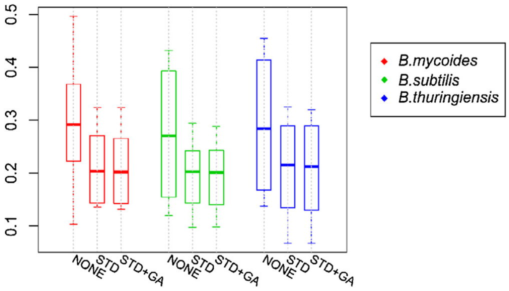


Fig. 4. Euclidean distances between the mean Raman spectra of three species were plotted in different colors. For each species, the distances were calculated for three cases: no spectral standardization, standard calibration, standard calibration plus wavenumber adjustment. The results are quite similar for different species. The distance was large, if no spectral standardization was applied, which decreased after the standard calibration. However, there was no further decrease by following wavenumber adjustment.

3.2. Parameter optimization

Besides the spectral standardization, the TR methods were investigated to update the current models. As it was mentioned before, an important task of TR methods is to optimize λ and η. To get an overview of the relationship between the model transferability and the parameters we firstly carried out a grid search based on the Raman spectra without spectral standardization. λ and η were changed exponentially on the basis of 10; the exponent was increased from -2 to 5 with a step size of 0.25 for λ, and from -5 to 5 with a step size of 0.5 for η. The matrices were constructed as aforementioned. To search optimal parameters an internal validation workflow was applied. Fifteen Raman spectra each five from one species were randomly chosen. These Raman spectra were termed as optimization sets. At each search step, we predicted the Raman spectra of the optimization sets of each device by seven models. Thus 28 accuracies were generated. The averaged value is saved as the results of the search step.

The overall results are plotted in logarithm coordinates in pseudo-color in Fig. 5. The first column shows the results of TR_1, while the other columns belong to the results of TR_2. In comparison, TR_2 is superior to TR_1. This results from the aforementioned instability of TR_1. Obviously, the model performance depends greatly on the value of λ. However, the exact value of η did not significantly affect the final results of TR_2, as long as η was a positive small value compared to the spectral intensity. Actually, this was predictable since η is just a numeric scaling in Eq. (4). If η was too large compared to the Raman spectral intensity, the multivariate calibration in Eq. (4) would be dominated by the identity matrix, decreasing the model transfer performance. Nevertheless, it is possible to simplify the two-parameter optimization by assigning η as a small positive value considering the large flat response of the model transferability to η. Additionally, this optimization can be sped up by heuristic optimization algorithms such as a genetic algorithm. This was achieved in our experiment with the logarithms of λ and η as genes. The averaged classification accuracy of the optimization sets

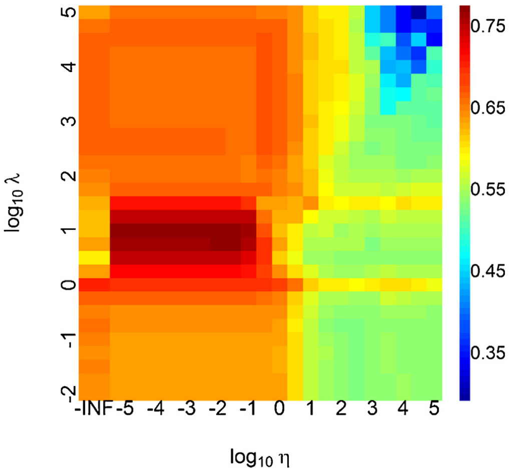


Fig. 5. Averaged accuracy of a classification model for different values of the parameters η and λ: The first column represents the result of TR_1. The following columns are the results of TR_2. It is seen that TR_1 performs worse than TR_2. Meanwhile, given a positive number below 10^{-1}, the exact value of η does not make a great difference to the final results. However, the accuracy decreases when η becomes larger than 10^{-1}. On the other hand, the accuracy greatly depends on the value of λ.

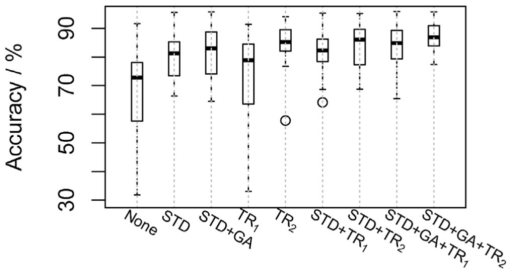


Fig. 6. Results of the prediction of Raman spectra measured on different spectrometers, with different model transfer approaches: ‘None’ on the abscissa means neither the spectral standardization nor the model updating was implemented. ‘STD’ represents standard calibration. The accuracy was disappointing if no model transfer procedure was applied, which is reflected by low mean values and high variations. This situation was improved by the standard calibration. By a following wavenumber adjustment, the model transferability was improved further. Additionally, the model transferability was improved by TR methods. However, the results of TR_1 were inadequate if no spectral standardization was performed. On the contrary, TR_2 performed always better than TR_1, given the same spectral standardization procedure. Comparable results as those after wavenumber adjustment were produced with TR_2 even without spectral standardization. Meanwhile, the performance of the TR_2 method was not greatly influenced by the conditional spectral variations. However, spectral standardization does improve the stability of TR_2.

was used as the evaluation function. The population size was set to 10 and the mutation chance was set to 0.75. The mutation rate was set high considering the fact that λ and η are not highly related. Therefore the crossover cannot ensure a fast convergence. The algorithm evolved for 30 generations before termination. Fig. S1 shows the values of λ, η, and the accuracy of the best chromosome within each generation in different cases of spectral standardization. The algorithm converged within 30 generations according to the value of the accuracy.

3.3. Classification performance

By now we have introduced two mechanisms of model transfer. For each mechanism, we have three alternatives: no spectral standardization, standard calibration, or the standard calibration plus the wavenumber adjustment for the spectral standardization. For the model updating we can use no model update, TR_1, or TR_2 method. In this section, we intend to evaluate the performance of these methods. The model transfer was achieved by either applying the two mechanisms separately or in combination. Accordingly, we studied nine combinations of model transfer procedures. As an evaluation, the accuracy of a three-class classification among the Raman spectra of the three spore species was calculated, as described in the section ‘Data Analysis’. No matter which model transfer method was applied, Raman spectra from each device were predicted by seven models, which was similar to the grid search mentioned above. Only Raman spectra, which were not used for the $\boldsymbol{L}$ matrix or as optimization sets were predicted. For the TR methods, the parameters λ and η were always optimized by the GA method. Accordingly, the optimal values of TR_1 were $\lambda = 10^{1.124}$, $10^{0.055}$, $10^{0.056}$ respectively for three above mentioned spectral standardization cases. $\eta = 10^{-4.034}$, $10^{-4.293}$, $10^{-4.197}$, $\lambda = 10^{0.701}$, $10^{1.020}$, $10^{0.598}$ were used for TR_2 correspondingly. The results of the four devices are plotted in Fig. 6. The model transfer methods were labeled on the abscissa. ‘None’ means neither the spectral standardization nor the model updating was implemented; while ‘STD’ represents the standard calibration. Detailed results for each combination were plotted in Fig. S2. With the results, several conclusions can be drawn.

Firstly, the accuracy was disappointing if no model transfer approach was used. Most accuracies were below 80%. It does not make a big difference, if the model is built based on data from multiple devices. The standard calibration greatly improved the accuracy: the lowest value was increased to 66.38%; the variation also was reduced. However, as was shown in Fig. S3, the accuracy may be decreased, for instance, the prediction of Raman spectra from device 3. This was possibly caused by the experimental deviations when measuring the standard materials on device 3, causing the poor performance of the standard calibration. Nevertheless, the accuracy was further improved by the wavenumber adjustment: accuracies were generally higher than the values after the standard calibration. The failure of standard calibration for the dataset of device 3 was corrected by the additional wavenumber adjustment. This shows the capability of the wavenumber adjustment for suppressing the influence of the experimental deviations during standard calibration, which leads to a higher stability of the model. Even though the Euclidean distances are only slightly reduced, a standard calibration in combination with an additional wavenumber adjustment significantly improved the model transferability.

Second, we performed the wavenumber alignment after the standard calibration in order to get a better wavenumber alignment. But the wavenumber alignment can also be used as a replacement of the standard calibration if the latter is not accessible. However, it is remarkable that the results of the wavenumber alignment greatly depend on the quality of the reference spectrum. If a standard calibration was not carried out beforehand, the model transferability would be not as good as it is shown in our experiment.

Third, the model transferability was improved by TR methods in all three cases of the spectral standardization, proven by higher accuracies than those of the model transfer without TR. Specifically, a large variation was observed for TR_1 without spectral standardization, which was probably caused by the abovementioned instability of TR_1. The results were much better if a spectral standardization was performed at first. On the other hand, TR$_2$ always performed better than TR$_1$, given the same spectral standardization procedure. Moreover, comparable results as those after wavenumber adjustment were produced with TR$_2$ even without spectral standardization. This is quite important especially if a standard calibration is not accessible. The performance of the TR$_2$ method is not greatly influenced by the change of spectral variations, at least in our experiment. Nevertheless, the variation of the accuracy decreased and the lower outlier margin increased by combining spectral standardization with TR$_2$, indicating an enhancement of model transferability.

To challenge the TR methods, only seven Raman spectra of each species were picked out from the secondary datasets during the model establishment, two for the ***L*** matrix and five for optimization sets. With such small number of new samples, we can already obtain quite satisfactory model transferability, especially with the TR$_2$ method. This indicates the great potential of TR$_2$ for model transfer problems, which is cheap and effective. Nonetheless, relatively low accuracies are still observed, which are probably caused by the sub-optimal values of λ and η. These results can be improved if more data are involved either for the ***L*** matrix or the optimization sets.

Finally, the accuracies were generally lower if ***X*** was constructed by Raman spectra of device 0 or 1 alone, which is shown in Fig. S2. A better performance was achieved by building models based on multiple devices, which means it is helpful to deal with a model transfer problem based on more than one primary datasets.

3.4. Model Stability

Besides the classification accuracy, we also evaluated the stability of the TR$_2$ method when using different Raman spectra for the ***L*** matrix. In detail, we classified Raman spectra of *B. mycoides* and *B. subtilis* measured by device 0, composed of three batches for each species. We did the experiment only with this two-class task, because only on device 0 were three batches of two species measured. ***X*** was constructed in a similar way as shown in Fig. 2, yet with only two species involved. We composed the ***L*** matrix with four Raman spectra, each two belonging to one species and picked randomly from any of the three batches. Thus nine ***L*** matrices were generated. Accordingly, we performed nine predictions for each ***X*** matrix. Parameters λ and η within TR$_2$ were optimized by the GA method every time with a new ***L*** matrix. The optimal values of the parameters are shown in Table S1. The accuracies are shown in the boxplots of Fig. 7. The device indices forming the ***X*** matrix were labeled on the abscissa. The two series correspond to the results with different Raman spectra of *B. subtilis* from batch 3 used for the ***L*** matrix. The correlation ($||s_1 - s_2||_2$) between the mean spectra of the selected Raman spectra of *B. subtilis* from different batches is shown in Fig. S3. For series 1, the variance of the classification accuracy is generally below 3%. An exception occurred if the ***X*** matrix was composed of the Raman spectra from device 1 alone, because three low accuracies were produced when employing the two Raman spectra of *B. subtilis* from batch 3 in the ***L*** matrix. As shown by the left plot of Fig. S3, these two Raman spectra demonstrated an obviously lower correlation compared with those from the other two batches. Therefore, we performed the classification again using other two Raman spectra of *B. subtilis* from batch 3 for the ***L*** matrix, which possessed a higher correlation (see the right side of Fig. S3). The results are plotted in series 2 in Fig. 7. Apparently, the variation of the accuracy greatly decreases. Thus a careful selection of the ***L*** matrix is helpful to ensure the stability of TR$_2$.

Nevertheless, the variation was not necessarily decreased by utilizing a better ***L*** matrix, comparing the two series of boxplots in Fig. 7. On the other hand, by re-checking the accuracy of the first boxplot, the variation of the accuracy was already dramatically decreased if Raman spectra from other devices or multiple devices were used for the ***X*** matrix. Thus the stability of TR$_2$ is more dependent on the ***X*** than the ***L*** matrix.

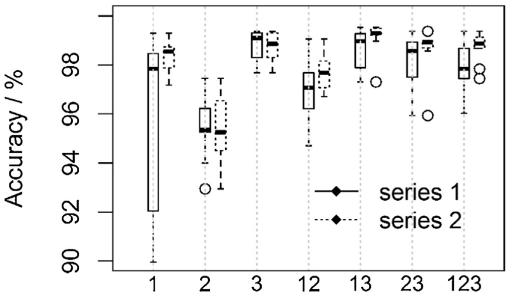


Fig. 7. Classification accuracy for Raman spectra of *B. mycoides* and *B. subtilis* measured on device 0: the device indices used to form the ***X*** matrix were labeled on the abscissa. Meanwhile, the ***L*** matrix was constructed by four Raman spectra from device 0, each two belonging to one species from any of the three batches. Thus nine possible ***L*** matrices were generated. Accordingly, nine classifications were performed for each ***X*** matrix, of which the accuracies were shown by boxplots. The two series correspond to results with different Raman spectra of *B. subtilis* from batch 3 selected for the ***L*** matrix. For series 1, the selected two Raman spectra of *B. subtilis* from batch 3 bear obvious variance compared to those from the other two batches (see Fig. S3). For series 2, we replaced these two Raman spectra with another two of *B. subtilis* from batch 3, featuring a higher correlation with the other two batches (see Fig. S3). Apparently, the variance of the accuracy is generally higher for series 1 than series 2, especially if ***X*** was composed from device 1 alone. Furthermore, the classification varied greatly with the ***X*** matrix, given the same ***L*** matrix. If the ***X*** matrix was composed of data from more than one device, the accuracy becomes much stable regardless of the goodness of the ***L*** matrix.

4. Conclusion

We reported an investigation to improve the model transferability for Raman spectroscopy in biological applications. Two mechanisms were studied including spectral standardization and model updating. For spectral standardization, we performed the wavenumber adjustment based on a genetic algorithm after the standard calibration. Current statistical models were updated by TR based methods. The different combinations of the two model transfer mechanisms were investigated and compared. To evaluate the spectral alignment, the Euclidean distances between mean Raman spectra of different devices were computed. To assess the model transferability, the classification accuracy of a three-class classification by PLSR was calculated. As was shown by the Euclidean distances, the standard calibration could dramatically decrease the between-device spectral differences. This improved the classification accuracy but the results are not adequate. By the following wavenumber alignment, the between-device spectral distances were almost the same but classification accuracy was observed to increase significantly. Meanwhile, the TR methods, especially the TR$_2$ method, yielded promising model transferability, even if no spectral standardization was carried out. However, a spectral standardization beforehand could help improve the stability of TR$_2$. Furthermore, the stability of TR$_2$ was strongly dependent on the ***X*** and less dependent on the ***L*** matrix.

With this investigation, we expect to step further in the model transfer problem for Raman spectroscopy. Despite of the demonstrated improvement, the results were not perfect. Actually, the accuracy of the statistical model was around 96%, if an independent dataset from the same device as the training set is predicted. This is better than the prediction of datasets from a different device to the training set after a model transform with the shown methods. Hence, more advanced model transfer approaches have to be developed. On the other hand, the model update was realized by TR in combination with a PLSR as a classifier. Other classifiers including linear discriminant analysis (LDA), support vector machine (SVM) and artificial neural networks (ANN) could also be potential alternatives to a PLSR. However, it is not straight forward to generalize the TR method to these classifiers, because the properties of the classifiers are different. For example, the parameter λ cannot be used for a LDA as for a PLSR. This is an open tissue and subject of our further investigations. Further research will also be related to the influence of the data size of the $\boldsymbol{L}$ matrix and optimization sets on the model transferability. Further research will be related to the influence of the data size of the $\boldsymbol{L}$ matrix and optimization sets on the model transferability.

Acknowledgements

The scholarship from China scholarship council (CSC) is greatly appreciated. Financial support of the BMBF and the EU via the project “PhotoSkin” (FKZ 13N13243) is highly appreciated. Funding of the projects “RamaDek” (FKZ 13N11168) and “InterSept” (FKZ 13N13852) by the BMBF is highly acknowledged.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.vibspec.2016.06.010.

References

[1] M. Diem, et al., Applications of infrared and Raman microspectroscopy of cells and tissue in medical diagnostics: present status and future promises, J. Spectrosc. 27 (5–6) (2012) 463–496.
[2] M. Diem, et al., Molecular pathology via IR and Raman spectral imaging, J. Biophotonics 6 (11–12) (2013) 855–886.
[3] C. Krafft, et al., Raman and coherent anti-Stokes Raman scattering microspectroscopy for biomedical applications, J. Biomed. Opt. 17 (4) (2012) 0408011–04080115.
[4] T.W. Bocklitz, et al., Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? Anal. Chem. 88 (1) (2016) 133–151.
[5] H. Abramczyk, B. Brozek-Pluska, Raman imaging in biochemical and biomedical applications. Diagnosis and treatment of breast cancer, Chem. Rev. 113 (8) (2013) 5766–5781.
[6] C. Bielecki, et al., Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells, J. Biomed. Opt. 17 (7) (2012) 0760301–0760308.
[7] W. Wang, et al., Real-time in vivo cancer diagnosis using raman spectroscopy, J. Biophotonics 8 (7) (2015) 527–545.
[8] E. Vargis, et al., Detecting biochemical changes in the rodent cervix during pregnancy using Raman spectroscopy, Ann. Biomed. Eng. 40 (8) (2012) 1814–1824.
[9] W. Richardson, et al., Ensemble multivariate analysis to improve identification of articular cartilage disease in noisy Raman spectra, J. Biophotonics 8 (7) (2015) 555–566.
[10] S.F. El-Mashtoly, et al., Label-free imaging of drug distribution and metabolism in colon cancer cells by Raman microscopy, Analyst 139 (5) (2014) 1155–1161.
[11] S. Pahlow, et al., Isolation and identification of bacteria by means of Raman spectroscopy, Adv. Drug Delivery Rev. 89 (2015) 105–120.
[12] S. Kloß, et al., Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study, Anal. Chem. 85 (20) (2013) 9610–9616.
[13] S. Stöckel, et al., Identification of Bacillus anthracis via Raman spectroscopy and chemometric approaches, Anal. Chem. 84 (22) (2012) 9873–9880.
[14] S. Stöckel, et al., Raman spectroscopic detection of anthrax endospores in powder samples, Angew. Chem. Int. Ed. 51 (22) (2012) 5339–5342.
[15] A. Walter, et al., Raman spectroscopic detection of Nickel impact on single Streptomyces cells-possible bioindicators for heavy metal contamination, J. Raman Spectrosc. 43 (8) (2012) 1058–1064.
[16] M. Jermyn, et al., Intraoperative brain cancer detection with Raman spectroscopy in humans, Sci. Trans. Med. 7 (274) (2015) 274ra19–274ra19.
[17] K. Virkler, I.K. Lednev, Blood species identification for forensic purposes using Raman spectroscopy combined with advanced statistical analysis, Anal. Chem. 81 (18) (2009) 7773–7777.
[18] N.K. Afseth, V.H. Segtnan, J.P. Wold, Raman spectra of biological samples: a study of preprocessing methods, Appl. Spectrosc. 60 (12) (2006) 1358–1367.
[19] M. Gregory, L.K. Igor, Spectroscopic discrimination of bone samples from various species, Am. J. Anal. Chem. 3 (2) (2012) 161–167.
[20] G. McLaughlin, K.C. Doty, I.K. Lednev, Discrimination of human and animal blood traces via Raman spectroscopy, Forensic Sci. Int. 238 (2014) 91–95.
[21] T. Fearn, Standardisation and calibration transfer for near infrared instruments: a review, J. Near Infrared Spectrosc. 9 (4) (2001) 229–244.
[22] P. Shahbazikhah, J.H. Kalivas, A consensus modeling approach to update a spectroscopic calibration, Chemom. Intell. Lab. Syst. 120 (2013) 142–153.
[23] J.H. Kalivas, et al., Calibration maintenance and transfer using Tikhonov regularization approaches, Appl. Spectrosc. 63 (7) (2009) 800–809.
[24] R.L. McCreery, Raman Spectroscopy For Chemical Analysis, vol. 157, John Wiley & Sons, 2000, 2016.
[25] T. Dörfer, et al., Checking and improving calibration of Raman spectra using chemometric approaches, Zeitschrift für Physikalische Chem. 225 (6–7) (2011) 753–764 (International journal of research in physical chemistry and chemical physics).
[26] Y. Wang, D.J. Veltkamp, B.R. Kowalski, Multivariate instrument standardization, Anal. Chem. 63 (23) (1991) 2750–2756.
[27] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Peak alignment using wavelet pattern matching and differential evolution, Talanta 83 (4) (2011) 1108–1117.
[28] J.H. Kalivas, Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, J. Chemom. 26 (6) (2012) 218–230.
[29] R.C., Team, R Foundation for Statistical Computing. Vienna, Austria, 3 (0) (2013).
[30] Signal developers, signal: signal processing (2013). http://r-forge.r-project.org/projects/.
[31] M., Morhac, Peaks: Peaks. R package version 0.2 (2012). http://CRAN.R-project.org/package=Peaks.
[32] K.H. Liland, B.-H. Mevik, baseline: Baseline Correction of Spectra. R package version 1.2-1 (2015). http://CRAN.R-project.org/package=baseline.
[33] T. Petzoldt, K. Rinke, Simecol: an object-oriented framework for ecological modeling in R, J. Stat. Software 22 (9) (2007) 1–31.
[34] W., Egon, B., Michel, genalg: R Based Genetic Algorithm. R package version 0.2.0 (2015). http://CRAN.R-project.org/package=genalg.
[35] B.-H., Mevik, R., Wehrens, K.H., Liland, pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0 (2015). http://CRAN.R-project.org/package=pls.
[36] J.D. Rodriguez, et al., Standardization of Raman spectra for transfer of spectral libraries across different instruments, Analyst 136 (20) (2011) 4232–4240.

Electronic Supplementary Material (ESI)

Towards an Improvement of Model Transferability for Raman Spectroscopy in Biological Applications

Shuxia Guo [a, b], Ralf Heinke [a, b], Stephan Stöckel [a, b], Petra Rösch [a, c], Thomas Bocklitz [a, b,*], Jürgen Popp [a, b, c]

a. *Institute of Physical Chemistry and Abbe School of Photonics, Friedrich-Schiller-University, Jena, Helmholtzweg 4, D-07743 Jena, Germany.*
b. *Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07745 Jena, Germany.*
c. *InfectoGnostics Research Campus Jena, Centre of Applied Research, Philosophenweg 7, D-07743, Jena, Germany.*

* Email:Thomas.bocklitz@uni-jena.de

Abstract

One of the most important issues for the application of Raman spectroscopy for biological diagnostics is how to deal efficiently with large datasets. The best solution is chemometrics, where statistical models are built based on a certain number of known samples and used to predict unknown datasets in future. However, the prediction may fail if the new datasets are measured under different conditions as those used for establishing the model. In this case, model transfer methods are required to obtain high prediction accuracy for both datasets. Known model transfer methods, for instance standard calibration and training models with datasets measured under multiple conditions, do not provide satisfactory results. Therefore, we studied two approaches to improve model transfer: wavenumber adjustment by a genetic algorithm (GA) after the standard calibration and model updating based on the Tikhonov regularization (TR). We based our investigation on Raman spectra of three spore species measured on four spectrometers. The methods were tested regarding two aspects. First, the wavenumber alignment is checked by computing Euclidean distances between the mean Raman spectra from different devices. Second, we evaluated the model transferability by means of the accuracy of a three-class classification system. According to the results, the model transferability was significantly improved by the wavenumber adjustment, even though the Euclidean distances were almost the same compared with those after the standard calibration. For the TR_2 method the model transferability was dramatically improved by updating current models with very few samples from the new datasets. This improvement was not significantly lowered even if no spectral standardization was implemented beforehand. Nevertheless, the model transferability was enhanced by combining different model transform mechanisms.

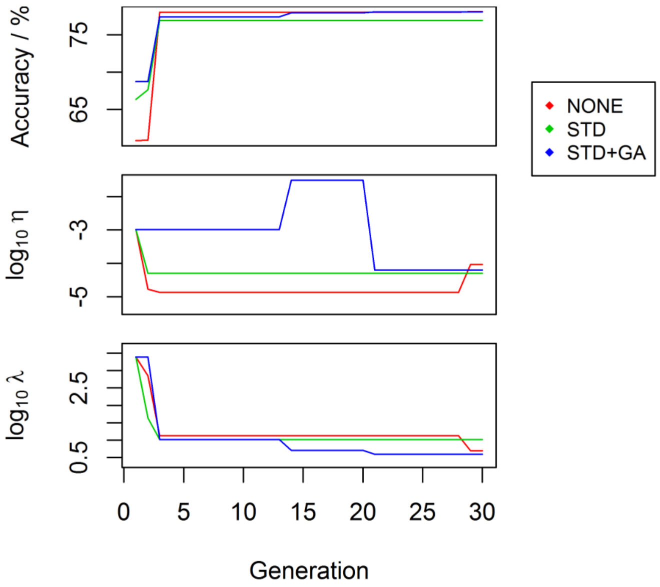


Figure S 1 Results of parameter optimization for TR_2 by genetic algorithm. The GA was carried out with population size of 10, evolving for 30 generations. The chromosome was composed of two genes, corresponding to the logarithm of the two parameters, λ and η. The gene values varied within [-2, 5] and [-5, 5] for λ and η, respectively. For each chromosome, 28 classifications were performed, in the same way as described for grid search. Their averaged accuracy was calculated as the evaluation function for GA, which was intended to be increased.

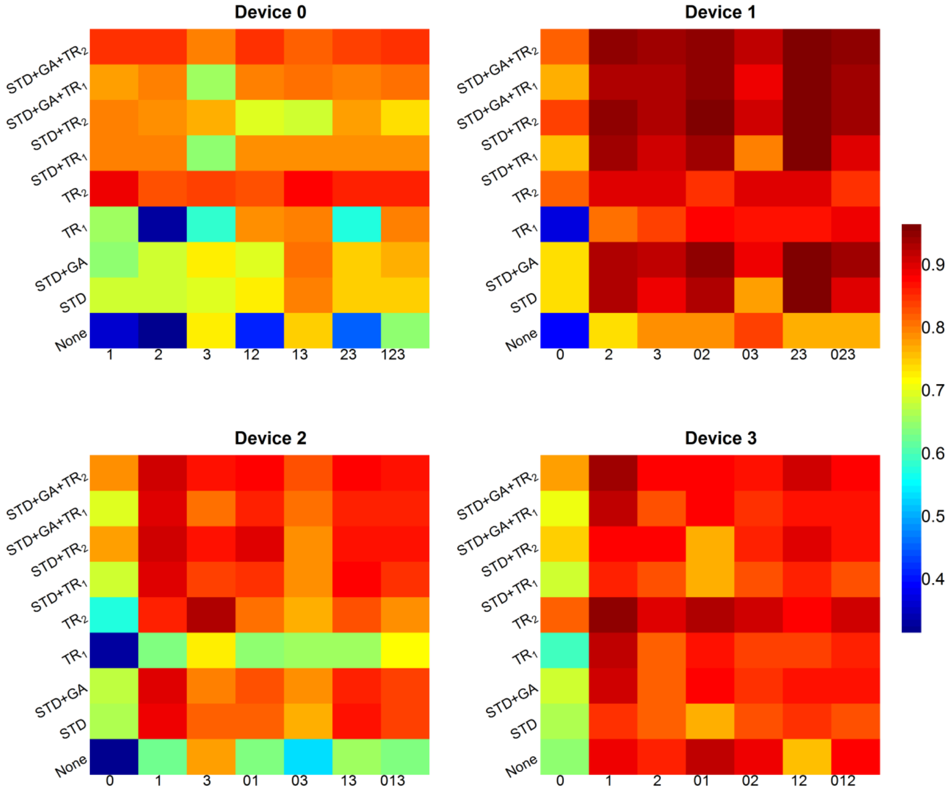


Figure S 2 False-color plots of classification accuracy for the three-class classification: different cases of model transfer were labeled on the ordinate, while the device indices of which the Raman spectra were used as training set were labeled on the abscissa. It was demonstrated that the accuracy was quite low if none model transfer approaches were used. The accuracy was not greatly increased by training model with Raman spectra measured on different devices, especially when predicting dataset measured by device 0 and 2. The standard calibration could improve the accuracy, which was yet not always successful. In some cases, for instance device 3, the accuracy declined after standard calibration. However, by further wavenumber adjustment, the accuracy could be generally increased. On the other hand, the TR_1 worked not as well as TR_2, given the same spectral standardization procedure. Furthermore TR_2 performed comparably when different spectral standardization progresses were carried out, all yielding promising results. Besides, the accuracy was quite low when predicting device 2 with TR_2 with a model trained by device 0. This was improved by performing spectral standardization before TR_2. It was also noteworthy that the accuracy was generally lower when only one device was used as training set. This indicated an advancement of model transform based on more than one primary datasets. Further improvement of TR can be expected by using more data during model establishment.

Table S 1 Optimal values of η and λ within TR$_2$ method for different $\boldsymbol{L}$ matrices. Subscripts of $\boldsymbol{L}$ represents the nine different combination of Raman spectra for composing $\boldsymbol{L}$ matrix. While η_1, λ_1, η_2, λ_2 refer to the parameters when different Raman spectra of batch 3 were used for $\boldsymbol{L}$.

	$\log_{10}(\eta_1)$	$\log_{10}(\lambda_1)$	$\log_{10}(\eta_2)$	$\log_{10}(\lambda_2)$
$\boldsymbol{L}_1$	-3.527	1.549	-3.527	1.549
$\boldsymbol{L}_2$	-2.983	0.722	-2.983	0.722
$\boldsymbol{L}_3$	-2.749	0.794	-4.767	2.849
$\boldsymbol{L}_4$	-0.352	1.344	-0.352	1.344
$\boldsymbol{L}_5$	-4.886	0.569	-4.886	0.569
$\boldsymbol{L}_6$	-2.749	0.641	-3.269	2.597
$\boldsymbol{L}_7$	-0.935	3.700	-0.935	3.700
$\boldsymbol{L}_8$	-4.490	0.687	-4.490	0.687
$\boldsymbol{L}_9$	-3.370	0.809	-2.983	2.211

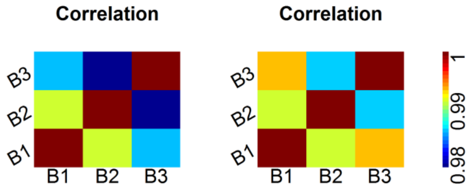


Figure S 3 Correlation of the mean spectra of the three groups of Raman spectra belonging to B. subtilis for constructing $\boldsymbol{L}$ matrix. For the first case, a big deviation was observed from batch 3 to other two batches, which is improved after using another two Raman spectra within batch 3.

7.5 Model Transfer for Raman Spectroscopy based Bacterial Classification (A5)

S. Guo, R. Heinke, S. Stöckel, P. Rösch, J. Popp, and T. Bocklitz, *Journal of Raman Spectroscopy*, 2018, 49, 627-637.

Der Nachdruck der folgenden Publikation erscheint mit freundlicher Genehmigung von Wiley. Reprinted with kind permission from Wiley.

Erklärungen zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation

Model transfer for Raman spectroscopy based bacterial classification, S. Guo[1], R. Heinke[2], S. Stöckel[3], P. Rösch[4], J. Popp[5], and T. Bocklitz[6], *Journal of Raman Spectroscopy*, 2018, 49, 627-637.						
Beteiligt an (*Zutreffendes ankreuzen*)						
	1	2	3	4	5	6
Konzeption des Forschungsansatzes	x				x	x
Planung der Untersuchungen	x	x	x	x	x	x
Datenerhebung		x	x	x		
Datenanalyse und -interpretation	x					x
Schreiben des Manuskripts	x			x	x	x
Vorschlag Anrechnung Publikationsäquivalente	1.0					

Received: 9 November 2017 | Revised: 9 January 2018 | Accepted: 15 January 2018
DOI: 10.1002/jrs.5343

RESEARCH ARTICLE

WILEY Journal of RAMAN SPECTROSCOPY

Model transfer for Raman-spectroscopy-based bacterial classification

Shuxia Guo[1,2] | Ralf Heinke[1] | Stephan Stöckel[1] | Petra Rösch[1] | Jürgen Popp[1,2,3] | Thomas Bocklitz[1,2]

[1] Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University of Jena, Helmholtzweg 4, D-07743 Jena, Germany
[2] Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07745 Jena, Germany
[3] InfectoGnostics Forschungscampus Jena, Philosophenweg 7, D-07743 Jena, Germany

Correspondence
Thomas Bocklitz, Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University of Jena, Helmholtzweg 4, D-07743 Jena, Germany.
Email: thomas.bocklitz@uni-jena.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Numbers: FKZ 13N13852 and FKZ 13N131243; Seventh Framework Programme, Grant/Award Numbers: FKZ 13N131243, FP 7 and CN 611682; China Scholarship Council, Grant/Award Number: for S.G.

Abstract
Raman spectroscopy has gained increasing attention in biomedical diagnostics thanks to instrumental developments and chemometric models that enhance the accuracy and speed of this technique. In particular, a model transfer procedure is needed if the chemometric models are utilized to predict a new dataset measured under (secondary) conditions different to the training data (primary). The model transfer methods try to achieve satisfactory prediction on the secondary dataset with minimal or no training samples measured under secondary conditions. Model transfer methods that have been reported are mostly applied for near-infrared spectroscopy and in regression problems. The investigation of model transfer in Raman spectroscopy and classification is rare. Our recently reported Tikhonov regularization based on partial least squares regression (TR-PLSR) was utilized for model transfer of Raman-based classification models for spore species. In the present work, we show that the TR-PLSR also works for Raman spectra of vegetative bacteria, even though the Raman spectra of 3 species of bacteria were acquired on 3 different Raman spectrometers. Additionally, we report 2 newly developed model transfer methods for Raman spectra: movement of principal components scores and spectral augmentation. Both methods were validated based on the Raman spectra of bacterial spores and vegetative bacteria, where a significant improvement of the model transferability was observed. The movement of principal components scores method yielded results comparable with those of the TR-PLSR. However, the new methods are superior to TR-PLSR in 2 ways: No training samples in the secondary conditions are necessary, and the methods are not restricted to partial least squares regression but can also be applied to other models. Both advantages are important in real-world applications and represent a large step for improving the model transfer of Raman spectra.

KEYWORDS
biological applications, classification, model transfer, Raman spectroscopy

Copyright © 2018 John Wiley & Sons, Ltd.

1 | INTRODUCTION

The field of chemometrics has developed dramatically within the past two decades, especially in biological diagnostics based on Raman spectroscopy. This is because chemometric methods can distinguish subtle Raman spectral differences related to biological changes, which is impossible by naked eye. The improvement of chemometrics for Raman spectral data led to two benefits: The sensitivity of biological diagnostics is significantly increased, and large-sized datasets can be analysed. These facts make the combination of Raman spectroscopy with chemometrics a versatile technique for biological applications,[1,2] including disease detection,[3–11] metabolic profiling,[12–14] bacterial identification,[15–18] intraoperative decision making,[19] and forensic analysis.[20,21]

Beyond the outlined biological application, it is well known in chemometrics that the prediction of an unknown dataset is mostly worse than the prediction of training data. This is known as shrinkage of predictors.[22] The shrinkage can be severer if the unknown data are measured under different conditions compared with the condition of the training data. The conditional changes, which manifest in Raman spectroscopy as wavenumber shifts and intensity variations, can be larger than the variations related to chemical or biological changes of interest. Consequently, the trained model completely fails to predict new data. The most straightforward solution is to construct a new model. However, this needs a large amount of new training samples and is not preferred. On the other hand, model transfer has emerged as an extremely important technique to improve the prediction of an existing model for new data with minimal or no requirement of new training samples measured under the changed conditions. Herein, the training dataset and the new dataset are termed as primary and secondary datasets, respectively. The aforementioned conditional changes could be experimental and instrumental changes or measurements of different individuals (e.g., patients) or biological replicates.

The existing model transfer approaches, including spectral standardization[23–26] and model updating,[27–29] were mostly developed for near-infrared spectroscopy and for regression models. Related investigations can be found in publications from the group of Professor Kalivas.[28–32] The model transfer for classification is rare[33] and often based on spectral standardization methods such as Procrustes analysis and piecewise direct standardization.[34] Particularly, orthogonal signal correction was applied to remove variations in the near-infrared spectra unrelated to the class property.[34] In addition, modified slope and bias correction was utilized by Myles et al.[35] to correct the difference between primary and secondary datasets in the score domain of a partial least squares model. However, both orthogonal signal correction and slope and bias correction required secondary data including metadata. This is often not possible in biological diagnosis, where the label of a new patient (e.g., the disease) needs to be predicted and is unknown. Spectral standardization of Raman spectroscopy is not trivial, especially in biological investigations. This is because the Raman spectra of biological samples are very complicated and spectral variations related to biological changes are tiny. Any artefacts introduced during the standardization might dramatically degrade the quality of further analysis.

Recently, model transfer for Raman-based classification was reported by Guo et al.,[36] where a Tikhonov-regularization-based partial least squares regression (TR-PLSR) method was developed. With TR-PLSR, the model was updated with a number of Raman spectra measured under secondary conditions. Despite the promising performance of TR-PLSR, it has two limitations. First, several secondary samples with known label information are required to transfer the model. Second, the TR-PLSR is developed for partial least squares regression (PLSR), which is a linear (regression) model. Its capability in nonlinear models is still an open issue. This limits its application for tasks where advanced classification or regression models are needed.

To tackle the aforementioned limitations, we report two new model transfer methods: movement of principal components scores (MS) and spectral augmentation (SA). The investigations were based on two datasets. The first dataset consists of Raman spectra of bacterial spore species (*Bacillus mycoides*, *Bacillus subtilis*, and *Bacillus thuringiensis*) measured on four spectrometers. The second dataset was composed of Raman spectra of vegetative bacteria of the same three species acquired on three spectrometers. As the Raman spectra of vegetative bacteria are normally more difficult to classify compared with those of spores, the vegetative bacterial dataset was used to further verify the TR-PLSR method. To evaluate the new model transfer methods, a three-class classifier was established based on a support vector machine (SVM) following a principal component analysis (PCA). The model transferability of the two new methods was compared with the standard wavenumber calibration and the TR-PLSR. The model transferability was benchmarked by the predictive performance of the models with respect to the secondary dataset.

2 | MATERIALS AND METHODS

2.1 | Spore cultivation and Raman spectroscopy

The information on cultivation and measurement of three spore species *B. mycoides* DSM 299, *B. thuringiensis* DSM

350, *B. subtilis* DSM 347, and *B. subtilis* DSM 10 strains can be found in our previous publications.[15,36] In summary, 1,592, 624, 654, and 848 Raman spectra were measured on four Raman spectrometers, almost equally distributed among the three bacteria species. The mean spectra of different species and the interspecies and interdevice Pearson correlation coefficients are visualized in Figure 1. Accordingly, the spectral variations between devices are larger than those between species. Moreover, the data measured on the first device are extremely dissimilar to those measured on the three other devices.

2.2 | Bacteria cultivation and Raman spectroscopy

The bacillus strains were cultivated on nutrient agar (peptone 5.0 g/L, meat extract 3.0 g/L, agar 15 g/L, distilled water 1,000 ml) for 24 hr at 30 °C. After cultivation, bacteria were washed three times with distilled water, and 10 μl of the bacteria solution was dropped on a nickel foil and dried.

The Raman spectra of a single bacterium on nickel foil were collected by a micro-Raman device (Bio Particle Explorer, rap.ID Particle Systems GmbH, Berlin, Germany) under ambient conditions. A 100× objective (MPLFLN 100xBD, Olympus Corporation, Tokyo, Japan) was used to focus the 532-nm excitation laser onto the sample with a spot diameter <1 μm. The laser power at the sample was 1 mW, and every Raman spectrum was integrated over 10 s. The 180° back-scattered Raman light was diffracted by a single-stage monochromator (HE 532, Horiba Jobin Yvon, Munich, Germany) with a 920 line/mm grating and collected with a thermoelectrically cooled charge-coupled device camera (DV401-BV, Andor Technology, Belfast, Northern Ireland) with a spectral resolution of ~8 cm^{-1}. The mean Raman spectra of different species and interspecies and interdevice Pearson correlation coefficients were visualized in Figure 2. In comparison with Figure 1, the Raman bands of spores are sharper than those of vegetative bacteria.[37] The standard deviations of the Raman spectra are smaller for the spores than for vegetative bacteria, demonstrating that spores are more reproducible than vegetative bacteria. Moreover, the Raman bands at 1,565, 1,440, 1,383, and 1,007 cm^{-1}, which

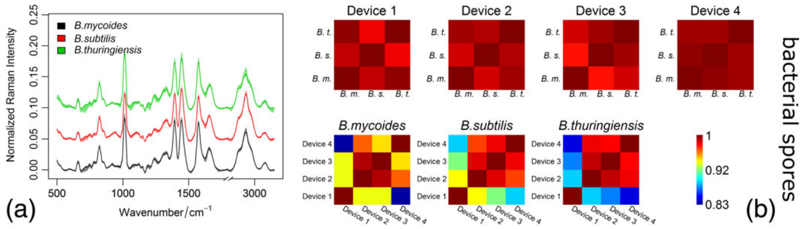


FIGURE 1 Overview of spore dataset. (a) Mean (solid line) and standard deviations (shade) of the Raman spectra measured on the first device. (b) Pearson correlation coefficients between the mean spectra of different species measured on the same device (first row) and those between the mean spectra of the same species measured on different devices (second row). Apparently, the spectral variations between devices are larger than those between species. Moreover, the data measured on the first device are extremely different to the data of the other three devices [Colour figure can be viewed at wileyonlinelibrary.com]

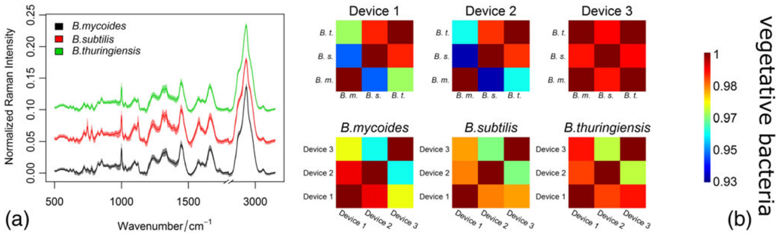
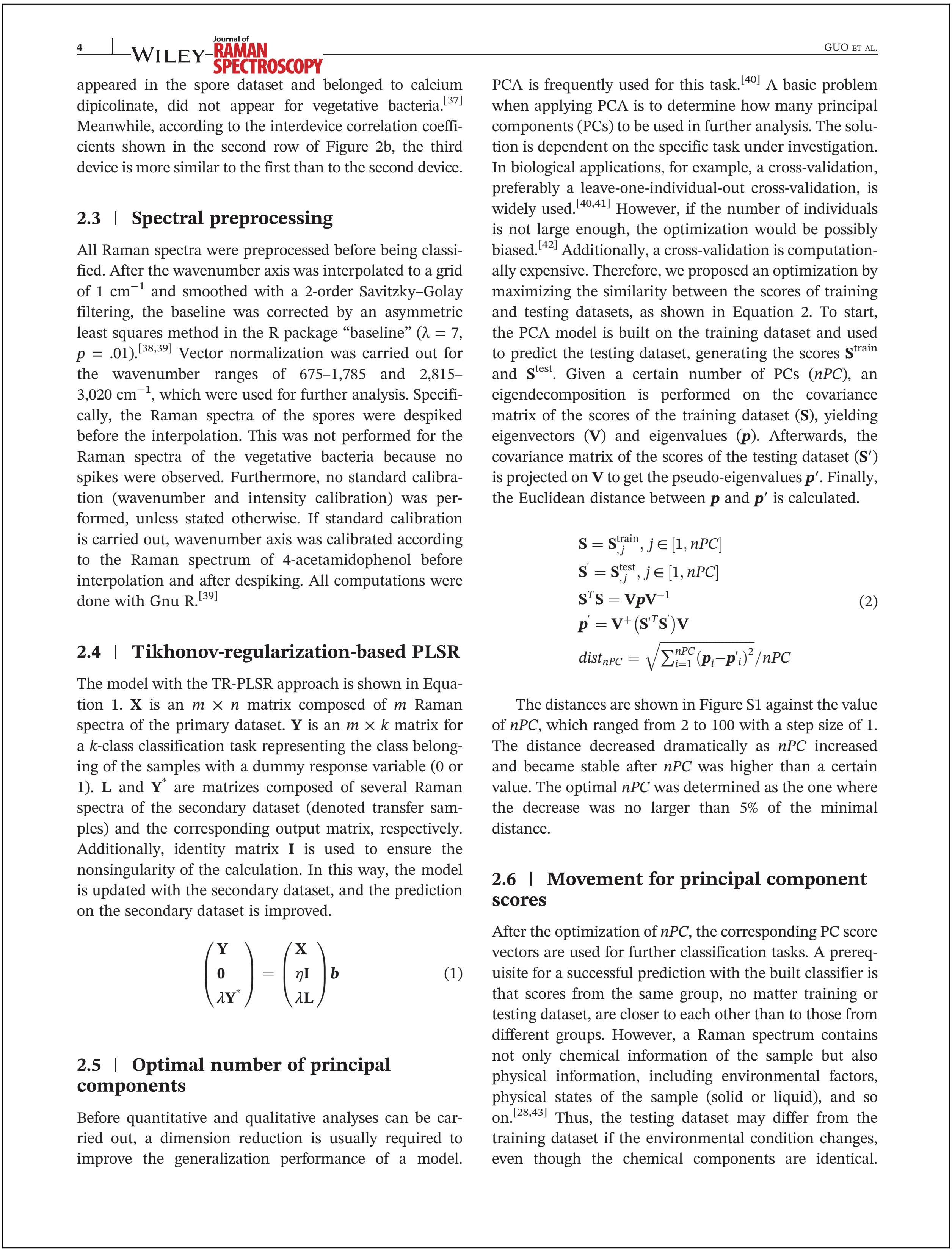


FIGURE 2 Overview of the bacterial dataset, visualized similarly as in Figure 1. Obviously, the Raman bands of vegetative bacteria are wider and less intense than the Raman bands of spores. Moreover, smaller standard deviations of the spore dataset represent a higher reproducibility compared with the vegetative bacteria. According to the interspecies correlation coefficients (first row in Plot b), the spectra measured on the first two devices are easier to separate than the spectra measured with the third device. Meanwhile, the interdevice correlation coefficients (second row in Plot b) demonstrated that the third device is more similar to the first than to the second device [Colour figure can be viewed at wileyonlinelibrary.com]



appeared in the spore dataset and belonged to calcium dipicolinate, did not appear for vegetative bacteria.[37] Meanwhile, according to the interdevice correlation coefficients shown in the second row of Figure 2b, the third device is more similar to the first than to the second device.

2.3 | Spectral preprocessing

All Raman spectra were preprocessed before being classified. After the wavenumber axis was interpolated to a grid of 1 cm^{-1} and smoothed with a 2-order Savitzky–Golay filtering, the baseline was corrected by an asymmetric least squares method in the R package “baseline” ($\lambda = 7$, $p = .01$).[38,39] Vector normalization was carried out for the wavenumber ranges of 675–1,785 and 2,815–3,020 cm^{-1}, which were used for further analysis. Specifically, the Raman spectra of the spores were despiked before the interpolation. This was not performed for the Raman spectra of the vegetative bacteria because no spikes were observed. Furthermore, no standard calibration (wavenumber and intensity calibration) was performed, unless stated otherwise. If standard calibration is carried out, wavenumber axis was calibrated according to the Raman spectrum of 4-acetamidophenol before interpolation and after despiking. All computations were done with Gnu R.[39]

2.4 | Tikhonov-regularization-based PLSR

The model with the TR-PLSR approach is shown in Equation 1. $\mathbf{X}$ is an $m \times n$ matrix composed of m Raman spectra of the primary dataset. $\mathbf{Y}$ is an $m \times k$ matrix for a k-class classification task representing the class belonging of the samples with a dummy response variable (0 or 1). $\mathbf{L}$ and $\mathbf{Y}^*$ are matrizes composed of several Raman spectra of the secondary dataset (denoted transfer samples) and the corresponding output matrix, respectively. Additionally, identity matrix $\mathbf{I}$ is used to ensure the nonsingularity of the calculation. In this way, the model is updated with the secondary dataset, and the prediction on the secondary dataset is improved.

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \\ \lambda \mathbf{Y}^* \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \eta \mathbf{I} \\ \lambda \mathbf{L} \end{pmatrix} \boldsymbol{b} \tag{1}$$

2.5 | Optimal number of principal components

Before quantitative and qualitative analyses can be carried out, a dimension reduction is usually required to improve the generalization performance of a model. PCA is frequently used for this task.[40] A basic problem when applying PCA is to determine how many principal components (PCs) to be used in further analysis. The solution is dependent on the specific task under investigation. In biological applications, for example, a cross-validation, preferably a leave-one-individual-out cross-validation, is widely used.[40,41] However, if the number of individuals is not large enough, the optimization would be possibly biased.[42] Additionally, a cross-validation is computationally expensive. Therefore, we proposed an optimization by maximizing the similarity between the scores of training and testing datasets, as shown in Equation 2. To start, the PCA model is built on the training dataset and used to predict the testing dataset, generating the scores $\mathbf{S}^{\text{train}}$ and $\mathbf{S}^{\text{test}}$. Given a certain number of PCs (nPC), an eigendecomposition is performed on the covariance matrix of the scores of the training dataset ($\mathbf{S}$), yielding eigenvectors ($\mathbf{V}$) and eigenvalues ($\boldsymbol{p}$). Afterwards, the covariance matrix of the scores of the testing dataset ($\mathbf{S}'$) is projected on $\mathbf{V}$ to get the pseudo-eigenvalues $\boldsymbol{p}'$. Finally, the Euclidean distance between $\boldsymbol{p}$ and $\boldsymbol{p}'$ is calculated.

$$\begin{aligned} &\mathbf{S} = \mathbf{S}^{\text{train}}_{,j},\ j \in [1, nPC] \\ &\mathbf{S}' = \mathbf{S}^{\text{test}}_{,j},\ j \in [1, nPC] \\ &\mathbf{S}^T\mathbf{S} = \mathbf{V}\boldsymbol{p}\mathbf{V}^{-1} \\ &\boldsymbol{p}' = \mathbf{V}^{+}\left(\mathbf{S}'^T\mathbf{S}'\right)\mathbf{V} \\ &dist_{nPC} = \sqrt{\textstyle\sum_{i=1}^{nPC}\left(\boldsymbol{p}_i - \boldsymbol{p}'_i\right)^2 / nPC} \end{aligned} \tag{2}$$

The distances are shown in Figure S1 against the value of nPC, which ranged from 2 to 100 with a step size of 1. The distance decreased dramatically as nPC increased and became stable after nPC was higher than a certain value. The optimal nPC was determined as the one where the decrease was no larger than 5% of the minimal distance.

2.6 | Movement for principal component scores

After the optimization of nPC, the corresponding PC score vectors are used for further classification tasks. A prerequisite for a successful prediction with the built classifier is that scores from the same group, no matter training or testing dataset, are closer to each other than to those from different groups. However, a Raman spectrum contains not only chemical information of the sample but also physical information, including environmental factors, physical states of the sample (solid or liquid), and so on.[28,43] Thus, the testing dataset may differ from the training dataset if the environmental condition changes, even though the chemical components are identical.

Therefore, the scores of the testing datasets are shifted compared with the scores of the training dataset, leading to a failed prediction of the classifier. This is severer if the inter-group differences are smaller compared with the spectral impact of the condition changes. A direct solution is to eliminate the spectral differences related to the physical information, for instance, by a spectral standardization. However, this is difficult for Raman-based biological applications because Raman spectral variations related to biological changes are very slight and can be corrupted by the artefacts possibly introduced by the standardization, which degrades further analysis.

$$\begin{aligned}
&\mathbf{X}^{\mathrm{pr}} = \mathbf{T}^{\mathrm{pr}}\mathbf{V}^{\mathrm{T}}, \mathbf{T}^{\mathrm{sc}} = \mathbf{X}^{\mathrm{sc}}\mathbf{V}\\
&\boldsymbol{T}^{\mathrm{mpr}} = \overline{\boldsymbol{X}}^{\mathrm{pr}}\mathbf{V}, \boldsymbol{T}^{\mathrm{mref}} = \overline{\boldsymbol{X}}^{\mathrm{ref}}\mathbf{V}\\
&\mathbf{T}^{\mathrm{tr}} = \mathbf{T}^{\mathrm{tr}} - \left(\boldsymbol{T}^{\mathrm{mpr}} - \boldsymbol{T}^{\mathrm{mref}}\right)
\end{aligned} \tag{3}$$

To avoid the drawback of spectral standardization, we corrected the shifts in the score space instead of the original spectral space. The scores of the training dataset were moved to match those of the testing dataset, which was done according to the differences between the scores of the mean spectra of the primary dataset and a reference dataset. The reference dataset was composed of spectra from the secondary dataset to be predicted. The calculation is shown in Equation 3, where $\mathbf{X}$, $\overline{\boldsymbol{X}}$, $\mathbf{T}$, and $\mathbf{V}$ represent the spectral matrix, mean spectrum, scores, and loadings, respectively. The superscripts pr, sc, and ref stand for the primary, secondary, and reference datasets, respectively. The additional prefix m represents the average spectrum of the corresponding datasets.

2.7 | Spectral augmentation

Whereas a model transfer procedure is achieved in the score space for MS method, the SA method aims to augment the original spectral space. The basic idea is to enlarge the spectral space of the training dataset, where the probability to successfully predict new data is increased. One of the possible options for such augmentation is to construct numerous spectra with the Raman spectra of pure substances composing the sample. However, there are a large number of pure components contained in a single biological sample, and this construction is hardly possible. Therefore, instead of simulating the chemical components, we imposed the spectral changes between the primary and secondary datasets resulting from physical (conditional) changes into the primary dataset. This was realized by enforcing certain wavenumber shifts and intensity variations into the primary dataset. First, the shift of each wavenumber (s_ν) was calculated from a fourth-order polynomial with randomly generated coefficients ($\boldsymbol{a}$) ranging from −1 to 1 ($s_\nu = a_1 + a_2\nu + a_3\nu^2 + a_4\nu^3 + a_5\nu^4$, $a_i \in [-1, 1]$). Second, we randomly sampled 20 wavenumber positions ($\min(\nu) < \nu_i^{\mathrm{smp}} < \max(\nu)$) and generated 20 random $\boldsymbol{b}$ values ranging from 0.5 to 2. The intensity response R at each wavenumber was obtained as a three-order polynomial ($R_{\nu_i} = c_1 + c_2\nu_i + c_3\nu_i^2 + c_4\nu_i^3$), where coefficients c were fitted from $\boldsymbol{b}$ and ν^{smp}. The intensity of the primary Raman spectrum was multiplied with $\boldsymbol{R}$.

Due to the randomness of the procedure, the resulting spectral changes may be unreasonably large, leading to invalid augmentation. To deal with this issue, we calculated the mean spectra from the primary and secondary datasets and located known Raman bands. Wavenumber shifts $\boldsymbol{d}$ and intensity ratios $\boldsymbol{r}$ of these Raman bands between the two mean spectra were calculated. Thereafter, the abovementioned wavenumber shift s_ν was enforced into range [max(quantile($\boldsymbol{d}$, 0.1), −5), min(quantile($\boldsymbol{d}$, 0.9), 5)] via a linear transformation. Similarly, an intensity response $\boldsymbol{R}$ was scaled to the range [max(quantile($\boldsymbol{r}$, 0.1), 0.5), min(quantile($\boldsymbol{r}$, 0.9), 2)]. The quantile calculation helps eliminate possible spectral outliers. Additional boundaries [−5, 5] and [0.5, 2] for s_ν and $\boldsymbol{R}$, respectively, are used to avoid values that are too large for $\boldsymbol{d}$ and $\boldsymbol{r}$. In this way, the augmentation was done by taking the differences between the primary and secondary datasets into consideration. The calculations were repeated for each spectrum of the primary dataset with different randomization of $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\nu}^{\mathrm{smp}}$. The group information was kept as it was. Unlike the model update, where the model was augmented with secondary samples, the proposed SA was done based on the primary dataset and does not require secondary samples during the model construction.

3 | RESULTS AND DISCUSSION

After introducing the theoretical background, in this section, we validate the model transfer methods with the two Raman spectral datasets described beforehand. The performance of the classifier was evaluated on the primary datasets by cross-validations, as described in the following sections. In particular, the k-fold cross-validation refers to splitting the training dataset randomly into k folds, and the distribution of different groups in a fold is kept the same as the overall training dataset. On the other hand, leave-one-batch-out cross-validation (LOBCV) refers to a splitting according to the information of biological replicates. By doing so, the data of different biological replicates are utilized as different folds. To check the model transfer methods, Raman spectra were split into primary and secondary datasets based on the devices they were measured on. A three-class classifier was built with an SVM (linear kernel, cost = 1) after a dimension reduction by PCA. The model transferability was estimated according to the predictions for the secondary datasets. In addition to MS and SA, the standard wavenumber calibration and the TR-PLSR were also performed for comparison. All predictions were benchmarked by the mean sensitivity of the three groups.

Before the results are presented, it is worth noting that besides an SVM with a linear kernel, other classification models could also be applied. We utilized an SVM with a linear kernel instead of a simpler linear discriminant analysis because linear discriminant analysis is more suitable for normal distributed data with comparable covariance matrices for different groups, which was hard to ensure in our investigation.

3.1 | Spore dataset

A PCA-SVM classification was performed on the spore Raman spectra measured by all devices with a 5-fold cross-validation. To do so, Raman spectra from the four devices were merged and randomly split into five folds for a cross-validation. Afterwards, the model transferability was tested, where the Raman spectra acquired by each device were used once as the secondary dataset. The primary dataset was composed of Raman spectra of all possible combinations of the other three devices ($C_3^1 + C_3^2 + C_3^3 = 7$ total possibilities). The mean sensitivity of each prediction on each secondary dataset was recorded, which resulted in 28 sensitivity values. The details are presented in Figure 3 and discussed in the following text.

Figure 3a visualised mean sensitivities in different cases of prediction. The result of the 5-fold cross-validation using Raman spectra from all four devices was shown in the black series, termed *Primary CV*, which demonstrated the satisfactory performance of the classifier. The predictions of the secondary datasets were shown in the other series, each including 28 values. Note that the suffix CV means the *nPC* was optimized by a 5-fold cross-validation on the training data; otherwise, the optimization method *optimal number of principal components* (OptNew) was used. The predictions on the secondary datasets without performing any model transfer methods were termed as *None CV* and *OptNew*, with mean sensitivities of 0.692 ± 0.180 and 0.687 ± 0.193, respectively. The results of the two *nPC* optimization methods were similar, indicating that they performed almost equally. However, the speed of OptNew was significantly enhanced because it did not require performing the classification for each *nPC* value.

To perform MS, the reference dataset was constructed with the first certain number of spectra of each species in the secondary dataset. The *nPC* was optimized by OptNew after the movement of the scores. The mean sensitivities were shown by the blue series in Figure 3a respective to the number of reference samples from each species (*MS without Group*). The values, for example, 0.830 ± 0.084, using 2 reference samples of each species, were significantly higher compared with those without model transfer procedures (0.692 ± 0.180). Furthermore, the performance of MS was almost independent of the number of spectra in the reference dataset.

Until now, we did not use the group information of the reference dataset, and the movement was done for all groups simultaneously. This suits applications where the group information of the secondary dataset is not accessible, such as for a diagnostic task. However, it is valuable to check if the group information can enhance the model transfer performance. To do so, we utilized the group information of the reference dataset and moved

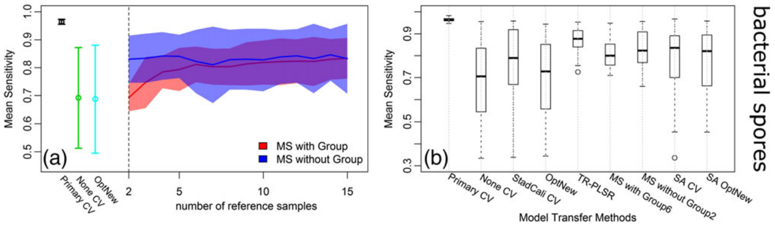


FIGURE 3 Results of the model transfer methods for the spore dataset. (a) The black series presents the prediction on all of the four devices with a five-fold cross-validation. The following two series represent the prediction of the secondary datasets without the model transfer procedure, where *nPC* was optimized with the two procedures described above. The other series represent the prediction with the movement of principal component scores (MS) method using varying numbers of spectra in the reference dataset. The blue and red series show the no-group and with-group MS, respectively. (b) Comparison of different model transfer methods. TR-PLSR and MS are comparable and superior to the standard wavenumber calibration (StdCali CV) and the spectral augmentation (SA) [Colour figure can be viewed at wileyonlinelibrary.com]

each group separately. The results were shown as the red series in Figure 3a (*MS with Group*). Apparently, the mean sensitivities were comparable with the sensitivities of the no-group MS if more than five reference samples of each group were utilized. Otherwise, the with-group MS was inferior and only provided mean sensitivities of 0.693 ± 0.049 if two reference samples of each species were used. This was because for the no-group MS, the mean spectrum was obtained from all spectra of the reference dataset, whereas this was computed from one group of the reference dataset for the with-group MS. Nonetheless, the group information is not necessary for the MS method, making it applicable for model transfer problems where the group information of the secondary dataset is unknown. This is advantageous to the bias and slope correction of Myles et al.,[35] where known secondary samples are required.

Afterwards, the SA method was tested with the same classifier. The PCA model was built on the original training dataset. The augmented training dataset and the testing dataset were predicted by this PCA model to receive their scores. The SVM was built on the scores of the augmented training dataset, with *nPC* ranging from 2 to 100. The predictions by the SA method and without any model transfer method are given in Figure S2. The predictions on the primary dataset based on a 5-fold cross-validation were plotted as a benchmark. For any value of *nPC*, the prediction on the secondary data was no better than the prediction on the primary dataset. Nonetheless, the prediction on the secondary dataset featured an obvious improvement with SA for model transfer compared with that without SA. For instance, given $nPC = 40$, the mean sensitivities with and without SA were 0.788 ± 0.110 and 0.696 ± 0.172, respectively. This demonstrated significantly enhanced model transferability by the SA method.

In addition to MS and SA, the standard wavenumber calibration and the TR-PLSR were performed as a comparison. It is worth noting that the two parameters of the TR-PLSR, λ and η, were optimized with a genetic algorithm by maximizing the mean sensitivity of the optimization samples instead of the accuracy as in a previous work.[36] The results were shown in Figure 3b. The standard wavenumber calibration led to a mean sensitivity of 0.773 ± 0.161, which was inferior to the other three model transfer methods for this dataset. The model transferability was comparable for the MS method and TR-PLSR method, which have mean sensitivities of 0.830 ± 0.084 and 0.869 ± 0.056, respectively. However, the MS method is superior in two ways. The group information of the reference dataset is not necessary for MS, and it can be applied to statistical models other than the PLSR. On the other hand, the mean sensitivities obtained with the SA method (0.779 ± 0.162) were lower than those obtained with MS and TR-PLSR. Nevertheless, this method provides a possibility of model transfer without knowledge of the secondary dataset as long as the ranges of the wavenumber shifts and intensity variations between two devices are roughly known.

Another important verification is to compare the prediction after model transfer with the local prediction by the model trained by the secondary dataset itself. To do so, we took the Raman spectra from the first device as the secondary dataset and performed classification with LBOCV. The sensitivities of the three species averaged to 0.864 and were plotted in the black series of Figure S3. The following series represented the results of the prediction if the Raman spectra from the other three devices were used as the primary (training) dataset. Apparently, the mean sensitivity was improved by all model transfer methods compared with the mean sensitivity without model transfer (0.484). The highest mean sensitivities of 0.806 and 0.787 were obtained with the TR-PLSR and no-group MS methods, respectively, which were yet inferior to the local prediction. This illustrated the limit of the model transfer methods and the requirement of further investigation.

3.2 | Bacterial dataset

To allow for a general conclusion, the model transfer methods were also verified by a second dataset, which contains the Raman spectra of three vegetative bacterial species collected by three Raman spectrometers. Specifically, one batch for each species was measured on the first two devices, whereas four batches for each species were measured on the third device. The Raman spectra from the third device were always used as the primary dataset, whereas the others were used as secondary datasets. The performance of PCA-SVM was verified on the primary dataset with LOBCV as well as 5-fold cross-validation. In terms of model transferability, five primary datasets were constructed with the Raman spectra of either all four batches or three out of the four batches measured on the third device.

Before the investigation of the MS and the SA, the TR-PLSR method was performed on this dataset, wherein the first 10 components were used for the PLSR. Two and five secondary spectra from each species were used as the reference dataset and optimization dataset, respectively. Parameters λ and η were optimized with a genetic algorithm by maximizing the mean sensitivity of the optimization dataset.[36] The results were presented in Figure S4. The mean sensitivities without TR-PLSR were 0.789 ± 0.120 and 0.698 ± 0.128 for the first and second devices, respectively, which were increased to 0.857 ± 0.009 and 0.857 ± 0.058, respectively, by the

TR-PLSR method. In particular, the predictions of the secondary datasets with TR-PLSR were even better than those of the primary dataset with LOBCV (0.765 ± 0.077). This could result from two factors. First, the primary dataset features small interspecies variances (first row in Figure 2b), whereas the interbatch variances are larger (Figure 4). Second, according to the first row of Figure 2b, the secondary datasets possess quite large interspecies differences compared with the primary dataset.

The MS and SA were carried out analogously to the datasets of the spores. The overall results for the two secondary devices were displayed in Figures 5. Detailed results were presented in Figures S5 and S6. Again, the OptNew gave similar results as a cross-validation for score determination. The prediction of the secondary dataset from Device 1 was slightly better than that for the primary dataset. This was due to the same factors described above for the results of TR-PLSR.

With respect to the MS method, the mean sensitivity of the prediction on the secondary dataset was improved from 0.778 ± 0.148 to 0.831 ± 0.075 by the no-group MS method using two reference samples of each species (Figure 5). The improvement was not as much as it was for the spore dataset, because the interdevice changes were smaller than those for the spore data, as shown in Figures 1b and 2b. However, the decreased variance represented better stability of the prediction with MS. Meanwhile, the no-group and with-group MS performed similarly if more than five reference samples were used. Nonetheless, as shown in Figure S4, the results of the with-group MS were inferior to those of the no-group MS for Device 1. This was because the utilized reference spectra were not good representatives of the corresponding group and their mean scores were largely biased from the centre of the whole group, as shown in Figure S7. Such bias can be decreased by the average of all groups together in the no-group MS. On the other hand, if such bias did not occur, the with-group MS could lead to better model transferability than the no-group MS, as for Device 2 in Figure S4. This told us that the no-group MS is superior to the with-group MS in terms of stability. A similar performance was observed for SA as for the spore dataset and thus is not further discussed; those results are given in Figure S8.

The results of the two devices predicted by the four model transfer methods were plotted in Figure 5b for comparison. The prediction with MS (0.831 ± 0.075) was comparable to the TR-PLSR (0.857 ± 0.039). The SA significantly improved the model transferability in this dataset from 0.778 ± 0.148 to 0.818 ± 0.111. In summary, a good transferability is possible without the requirement of label information from the secondary dataset with the

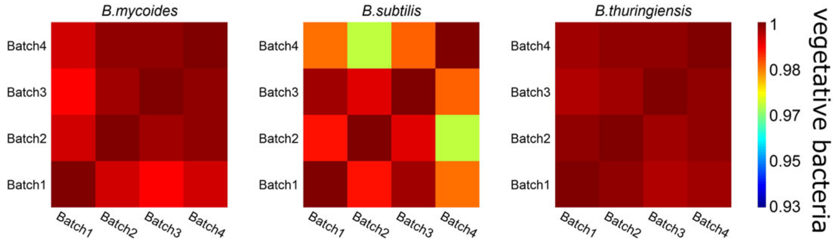


FIGURE 4 Pearson correlation coefficients between the mean spectra of different batches measured on the third spectrometer belonging to the same vegetative bacterial species. Compared with those in the first row in Figure 2b, the interbatch variations are larger than interspecies differences [Colour figure can be viewed at wileyonlinelibrary.com]

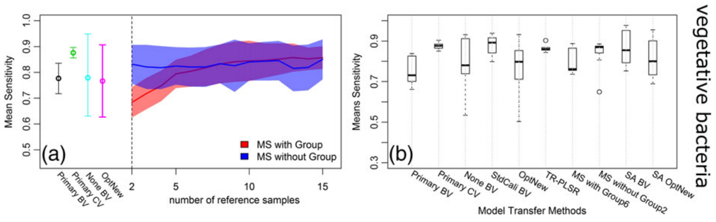


FIGURE 5 Results of the model transfer methods on vegetative bacterial dataset, plotted similarly as in Figure 3. Particularly, the performance of the classifier was verified with both a normal 5-fold cross-validation (primary CV) and a leave-one-batch-out cross-validation (primary BV) on the primary dataset [Colour figure can be viewed at wileyonlinelibrary.com]

MS and SA methods. In addition, it is clear that the MS method can be generalized to other factor analysis methods such as partial least squares, independent component analysis, and so on. The capability of SA is limited due to its random procedure.

3.3 | Stability test for MS

By now, we already checked the capability of MS for model transfer problems. In this section, we test its stability in two aspects: repeated samplings for the reference dataset and the number of groups included in the reference dataset. The validation was done based on the Raman spectra of the spores. Raman spectra measured on the first device were used as the secondary datasets, because we have three batches of species *B. mycoides* and *B. subtilis* for this device, which makes the repeated sampling more meaningful. Seven possible combinations of the other three devices were used as primary datasets. Both the no-group and with-group MS methods were investigated.

First, to check the influence of the number of reference spectra, we randomly selected a certain number of the Raman spectra from the first device as reference spectra. The number of samples taken from each species was the same and ranged from 2 to 15 with a step of one. For each reference spectra number, the sampling was repeated 100 times. After each sampling, the MS was carried out to predict the secondary dataset with the model trained on the seven primary datasets. The resulting seven mean sensitivities were averaged. The 100 repetitions resulted in 100 values for each reference spectra number, as shown in Figure 6b. The mean sensitivities of the prediction on the primary and secondary datasets without model transfer were given in Figure 6a. Both the no-group and with-group MS improved the model transferability dramatically with high stability. The with-group MS was statistically superior to the no-group MS according to the higher mean sensitivities. However, the with-group MS featured an inferior stability shown by the larger standard deviations over different samplings.

Second, we investigated the influence of the number of groups included in the reference dataset. To do so, the reference dataset was sampled from different numbers of groups (one, two, or three). Additionally, the overall number of reference spectra was varied among 5, 10, 15, and 20. We repeated the sampling 50 times for each possible combination of number of groups and spectra and performed both no-group and with-group MS. For the with-group MS, the scores of the groups not included in the reference dataset were moved according to the averaged movement of all reference spectra. The results of no-group and with-group MS were presented in Figure 6c as boxes without and with coloured fillings, respectively. Generally, model transferability was significantly improved regardless of the number of groups or the number of spectra included in the reference dataset. However, the number of groups did influence the performance, especially for the no-group MS. In this case, the performance of model transfer dramatically decreased if the reference spectra were from only one group. Nevertheless, the performance can be improved by increasing

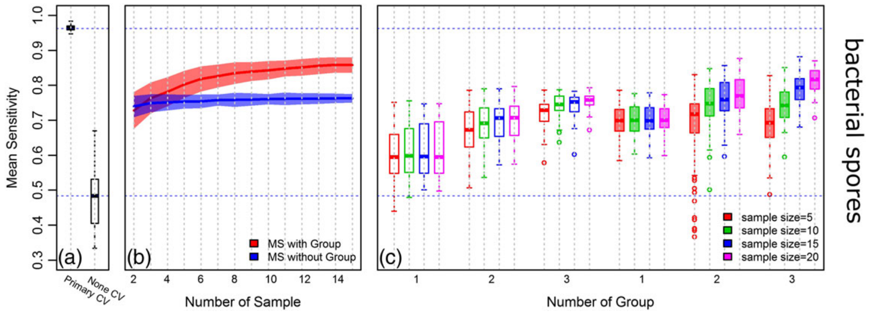


FIGURE 6 Results of the stability test for movement of principal components scores (MS) on the spore dataset. (a) The prediction of the primary datasets and the secondary datasets without model transfer procedure. (b) Results of the 100 repeated samplings, with different number of spectra used in the reference datasets. Both no-group and with-group MS methods demonstrate high stability. The with-group MS is superior to the no-group MS if more than five spectra are selected from each species as the reference dataset. (c) Results of MS with different numbers of groups included in the reference datasets. The results of no-group and with-group MS are plotted as boxes without or with fillings, respectively. Different numbers (5, 10, 15, and 20) of spectra were sampled as the reference dataset. The model transferability is significantly improved by MS, regardless of the number of groups or the number of spectra in the reference dataset. However, the model transfer performance is dependent on the number of groups; for instance, the no-group MS performs worse if the reference dataset is composed of spectra from one single group only. Nevertheless, the model transfer performance can be improved if more spectra are used as a reference [Colour figure can be viewed at wileyonlinelibrary.com]

the number of reference spectra. In principle, all testing spectra can be used as references because no label information is required. Third, the performance of with-group MS was independent of the number of groups in the reference dataset, as long as more than five spectra were used in the reference dataset.

Even though the results shown above are quite promising, there is one important issue to be kept in mind regarding the MS method. The distribution of the PC scores is assumed to not change significantly between the primary and secondary datasets. If this is violated, a with-group MS is preferred to the no-group MS.

3.4 | Method comparison

The model transfer performance was compared among the abovementioned methods on the basis of the results of prediction on the secondary datasets shown in Figures 3 and 5. The sum of ranking differences was utilized, which is capable of achieving a fair model comparison.[44] More details and the results are given in Figure S9, which gives conclusions consistent with those of Figures 3 and 5.

4 | CONCLUSION

We reported about two new model transfer methods (MS and SA) to deal with model transfer problems in Raman spectroscopy, especially if label information of the secondary dataset is absent. Their performance was benchmarked by a three-group bacterial classification task and compared with the recently published TR-PLSR approach for model transfer. All carried out validations showed that both methods were able to significantly improve model transferability. For the dataset from bacterial spores, the mean sensitivity of the prediction on the secondary dataset was improved from 0.692 ± 0.180 to 0.830 ± 0.084 with the no-group MS and to 0.779 ± 0.162 with SA. For the dataset from vegetative bacteria, the mean sensitivity of the prediction on the secondary dataset was improved from 0.778 ± 0.148 to 0.831 ± 0.075 with the no-group MS and to 0.818 ± 0.111 with SA. The performance was slightly inferior but comparable with TR-PLSR. However, the new methods are superior to TR-PLSR in two aspects: First, MS and SA do not need group information of the secondary datasets, which is extremely important for biological applications where the group information of the new samples (secondary dataset) is unknown, for example, in medical diagnosis. Second, the proposed methods are not limited to certain classification or regression models, whereas TR-PLSR being limited to PLSR. This makes it possible to tackle the model transfer problem for advanced statistical models, which are often required in biological applications of Raman spectroscopy. In addition, MS features the potential to be generalized to other factor analysis methods such as nonnegative matrix factorization or independent component analysis. However, the methods are still limited because they are designed specifically for classification tasks and are not applicable for regression methods. Investigating advanced approaches to tackle this limitation is the topic of future research.

ACKNOWLEDGEMENTS

The scholarship from the China Scholarship Council (CSC) for S. G. is greatly appreciated. Financial support by the Bundesministerium für Bildung und Forschung (BMBF) and the European Union (EU) via the project “PhotoSkin” (FKZ 13N131243), the support by the EU for “HemoSpec” (Seventh Framework Programme, CN 611682) and funding of the BMBF for “Intersept” (FKZ 13N13852) are highly acknowledged.

ORCID

Thomas Bocklitz http://orcid.org/0000-0003-2778-6624

REFERENCES

[1] T. W. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler, J. Popp, *Anal. Chem.* **2016**, *88*(1), 133.

[2] A. C. S. Talari, Z. Movasaghi, S. Rehman, I. U. Rehman, *Appl. Spectrosc. Rev.* **2015**, *50*(1), 46.

[3] E. M. Barroso, R. W. H. Smits, T. C. Bakker Schut, I. ten Hove, J. A. Hardillo, E. B. Wolvius, R. J. Baatenburg de Jong, S. Koljenović, G. J. Puppels, *Anal. Chem.* **2015**, *87*(4), 2419.

[4] K. Kong, C. Kendall, N. Stone, I. Notingher, *Adv. Drug Delivery Rev.* **2015**, *89*, 121.

[5] P. Matousek, N. Stone, *Chem. Soc. Rev.* **2016**, *45*(7), 1794.

[6] A. Pallaoro, M. R. Hoonejani, G. B. Braun, C. D. Meinhart, M. Moskovits, *ACS Nano* **2015**, *9*(4), 4328.

[7] I. P. Santos, P. J. Caspers, T. C. Bakker Schut, R. van Doorn, V. Noordhoek Hegt, S. Koljenović, G. J. Puppels, *Anal. Chem.* **2016**, *88*(15), 7683.

[8] W. Wang, J. Zhao, M. Short, H. Zeng, *J. Biophotonics* **2015**, *8*(7), 527.

[9] T. Meyer, N. Bergner, C. Krafft, D. Akimov, B. Dietzek, J. Popp, C. Bielecki, B. F. M. Romeike, R. Reichart, R. Kalff, *J. Biomed. Opt.* **2011**, *16*(2). 021113-021113-10

[10] N. Vogler, T. Bocklitz, F. S. Salah, C. Schmidt, R. Bräuer, T. Cui, M. Mireskandari, F. R. Greten, M. Schmitt, A. Stallmach, I. Petersen, J. Popp, *J. Biophotonics* **2016**, *9*(5), 533.

[11] N. Stone, C. Kendall, J. Smith, P. Crow, H. Barr, *Faraday Discuss.* **2004**, *126*, 141.

[12] B. Berry, J. Moretto, T. Matthews, J. Smelko, K. Wiltberger, *Biotechnol. Prog.* **2015**, *31*(2), 566.

[13] S. F. El-Mashtoly, D. Petersen, H. K. Yosef, A. Mosig, A. Reinacher-Schick, C. Kötting, K. Gerwert, *Analyst* **2014**, *139*(5), 1155.

[14] V. Kumar BN, S. Guo, T. Bocklitz, P. Rösch, J. Popp, *Anal. Chem.* **2016**, *88*(15), 7574.

[15] S. Stöckel, S. Meisel, M. Elschner, P. Rösch, J. Popp, *Anal. Chem.* **2012**, *84*(22), 9873.

[16] D. Berry, K. B. Mahfoudh, M. Wagner, A. Loy, *Appl. Environ. Microbiol.* **2011**, *77*(21), 7846.

[17] A. Walter, S. Kuhri, M. Reinicke, T. Bocklitz, W. Schumacher, P. Rösch, D. Merten, G. Büchel, E. Kothe, J. Popp, *J. Raman Spectrosc.* **2012**, *43*(8), 1058.

[18] B. Lorenz, C. Wichmann, S. Stöckel, P. Rösch, J. Popp, *Trends Microbiol.* **2017**, *25*(5), 413.

[19] M. Jermyn, K. Mok, J. Mercier, J. Desroches, J. Pichette, K. Saint-Arnaud, L. Bernstein, M. C. Guiot, K. Petrecca, F. Leblond, *Sci. Transl. Med.* **2015**, *7*(274), 274ra19.

[20] P. Buzzini, E. Suzuki, *J. Raman Spectrosc.* **2016**, *47*(1), 16.

[21] K. Virkler, I. K. Lednev, *Anal. Bioanal. Chem.* **2010**, *396*(1), 525.

[22] J. B. Copas, *J. R. Stat. Soc. Ser. B Methodol.* **1983**, *45*(3), 311.

[23] T. Bocklitz, T. Dörfer, R. Heinke, M. Schmitt, J. Popp, *Spectrochim. Acta, Part A* **2015**, *149*, 544.

[24] T. B. Blank, S. T. Sum, S. D. Brown, S. L. Monfre, *Anal. Chem.* **1996**, *68*(17), 2987.

[25] T. Dörfer, T. Bocklitz, N. Tarcea, M. Schmitt, J. Popp, *Zeitschrift für Physikalische Chemie Int. J. Res. Phys. Chem. Chem. Phys.* **2011**, *225*(6–7), 753.

[26] C. Liang, H. F. Yuan, Z. Zhao, C. F. Song, J. J. Wang, *Chemom. Intell. Lab. Syst.* **2016**, *153*, 51.

[27] T. Fearn, *J. Near Infrared Spectrosc.* **2001**, *9*(4), 229.

[28] J. H. Kalivas, G. G. Siano, E. Andries, H. C. Goicoechea, *Appl. Spectrosc.* **2009**, *63*(7), 800.

[29] J. Ottaway, J. H. Kalivas, *Appl. Spectrosc.* **2015**, *69*(3), 407.

[30] J. A. Farrell, K. Higgins, J. H. Kalivas, *J. Pharm. Biomed. Anal.* **2012**, *61*, 114.

[31] J. H. Kalivas, *J. Chemom.* **2012**, *26*(6), 218.

[32] J. H. Kalivas, J. Palmer, *J. Chemom.* **2014**, *28*(5), 347.

[33] M. Blackburn, S. Ramos, B. Rohrback. Transfer of calibration for classification problems. InfoMetrix **2002**.

[34] J. A. F. Pierna, A. B. Sanfeliu, B. Slowikowski, C. von Holst, O. Maute, L. Han, G. Amato, B. de la Roza Delgado, D. P. Marín, G. Lilley, P. Dardenne, V. Baeten, *Biotechnol. Agron. Soc. Environ.* **2013**, *17*(4), 547.

[35] A. J. Myles, T. A. Zimmerman, S. D. Brown, *Appl. Spectrosc.* **2006**, *60*(10), 1198.

[36] S. Guo, R. Heinke, S. Stöckel, P. Rösch, T. Bocklitz, J. Popp, *Vib. Spectrosc.* **2017**, *91*, 111.

[37] P. Rösch, M. Harz, M. Schmitt, K. D. Peschke, O. Ronneberger, H. Burkhardt, H.-W. Motzkus, M. Lankers, S. Hofer, H. Thiele, J. Popp, *Appl. Environ. Microbiol.* **2005**, *71*(3), 1626.

[38] K. H. Liland, B.-H. Mevik, Baseline: baseline correction of spectra **2015**.

[39] R. C. Team, *R Foundation for Statistical Computing*, Vienna, Austria **2013**.

[40] R. Bro, A. K. Smilde, *Anal. Methods* **2014**, *6*(9), 2812.

[41] S. Arlot, A. Celisse, *Stat. Surv.* **2010**, *4*, 40.

[42] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, J. Popp, *Anal. Chim. Acta* **2013**, *760*, 25.

[43] S. Pahlow, S. Meisel, D. Cialla-May, K. Weber, P. Rösch, J. Popp, *Adv. Drug Delivery Rev.* **2015**, *89*, 105.

[44] K. Héberger, *TrAC Trends Anal. Chem.* **2010**, *9*(1), 101.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Guo S, Heinke R, Stöckel S, Rösch P, Popp J, Bocklitz T. Model transfer for Raman-spectroscopy-based bacterial classification. *J Raman Spectrosc*. 2018;1–11. https://doi.org/10.1002/jrs.5343

Electronic Supplementary Material (ESI)

Model transfer for Raman spectroscopy based bacterial classification

Shuxia Guo[a,b], Ralf Heinke[a], Stephan Stöckel[a], Petra Rösch[a], Jürgen Popp[a,b,c] and Thomas Bocklitz*[a,b]

[a]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University of Jena, Helmholtzweg 4, D-07743 Jena, Germany

[b]Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, D-07745 Jena, Germany

[c] InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, D-07743 Jena, Germany

ABSTRACT: Raman spectroscopy has gained increasing attention in biomedical diagnostics thanks to instrumental developments and chemometric models that enhance the accuracy and speed of this technique. In particular, model transfer procedure is needed if the chemometric models are utilized to predict new dataset measured under (secondary) conditions different to the training data (primary). The model transfer methods try to achieve satisfactory prediction on the secondary dataset with minimal or no training samples measured under secondary conditions. Model transfer methods that have been reported are mostly applied for near-infrared spectroscopy and in regression problems. The investigation of model transfer in Raman spectroscopy and classification is rare. Our recently reported Tikhonov regularization based on partial least squares regression (TR-PLSR) was utilized for model transfer of Raman based classification models for spore species. In the present work, we show that the TR-PLSR also works for Raman spectra of vegetative bacteria, even though the Raman spectra of three species of bacteria were acquired on three different Raman spectrometers. Additionally, we report two newly developed model transfer methods for Raman spectra: movement of principal components scores (MS) and spectral augmentation (SA). Both methods were validated based on the Raman spectra of bacterial spores and vegetative bacteria, where a significant improvement of the model transferability was observed. The MS method yielded comparable results to the TR-PLSR. However, the new methods are superior to TR-PLSR in two ways: No training samples in the secondary conditions are necessary and the methods are not restricted to PLSR but can also be applied to other models. Both advantages are important in real-world applications and represent a large step for improving the model transfer of Raman spectra.

Keywords: model transfer; Raman spectroscopy; biological applications; classification;

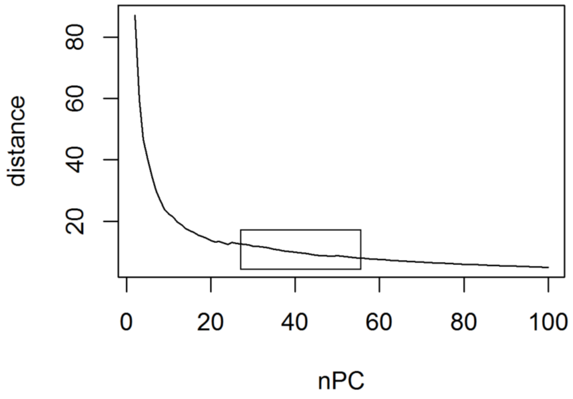


Figure S1 Euclidean distance between the clusters of PC scores from training and testing sets varying with the value of nPC.

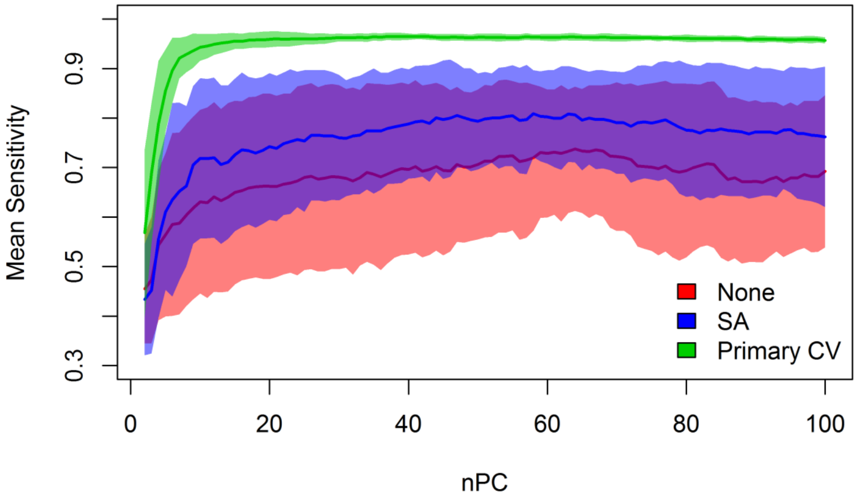


Figure S2 Prediction of spore dataset with varied nPC values: the green series shows the results of 5-fold cross-validation on all the Raman spectra measured on the four devices. The red series present the results of prediction on the secondary sets without model transform. The blue series demonstrates the results of prediction on the secondary sets using SA for a model transform.

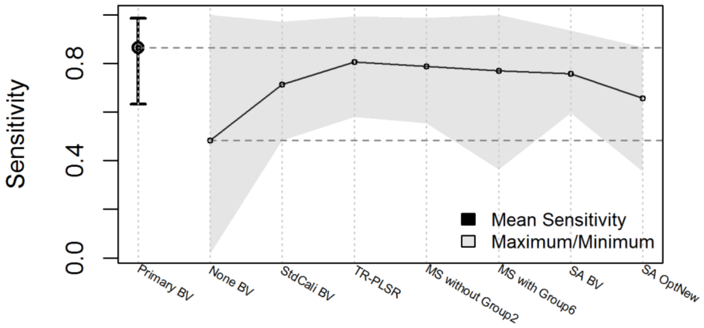


Figure S3: The sensitivities of the three species from the local prediction (first column) and the prediction with model transfer (the other columns), where the Raman spectra from the first and the other three devices were used as the secondary and the primary dataset, respectively. The mean, maximum, and the minimum of the three sensitivities were visualized for each case of prediction. The local prediction was achieved with a leave-one-batch-out cross-validation on the secondary dataset, of which the results indicated the best classification can be possibly achieved. Apparently, the model transfer method could significantly improve the prediction from the primary dataset to the secondary datasets. However, none of the methods could achieve the 'best' classification.

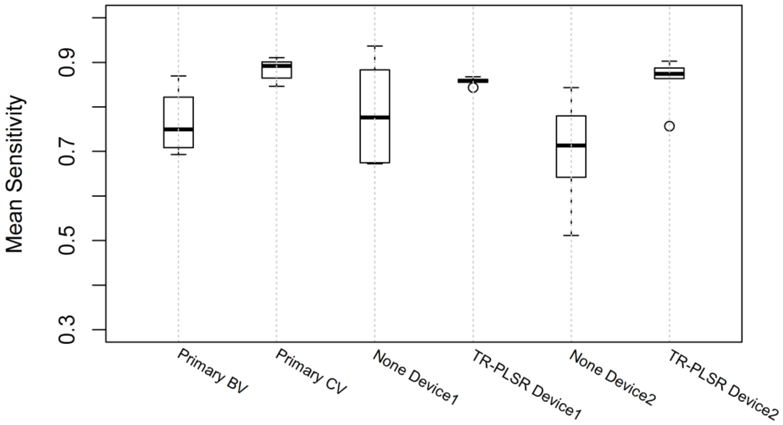


Figure S4 Results of PLSR on the bacterial datasets. The first two columns are prediction of the spectra measured on the third device, used as the primary set, with a LOBCV and a normal 5-fold cross-validation. The following columns show the prediction on the secondary set, i.e., spectra measured on the first two devices. 'none' means results with no model transform, while 'TR' means prediction using TR-PLSR as a model transform.

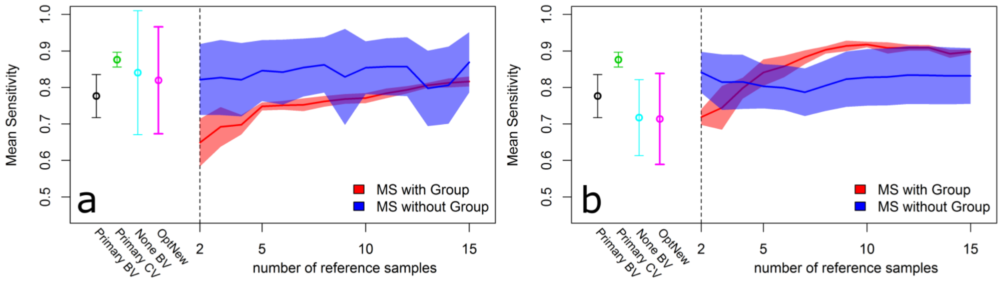


Figure S5 the similar results as shown in Figure 2c. Here the results of bacterial Raman spectra measured on the first and second devices are plotted separately in a and b, respectively.

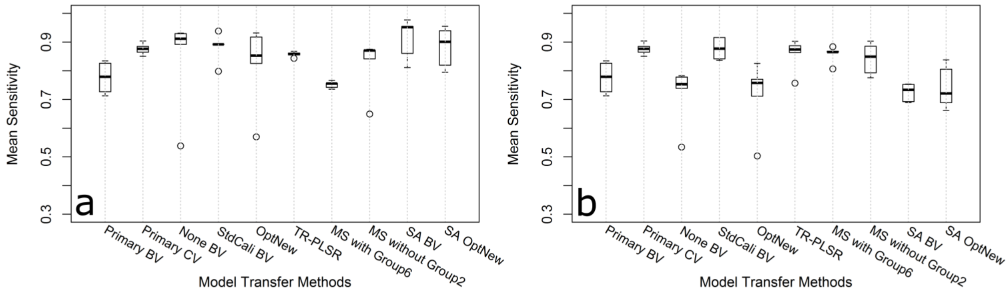


Figure S6 the similar results as shown in Figure S2, from the bacterial datasets. The prediction of the primary set (in green) was obtained by a LOBCV instead of a normal 5-fold cross validation as in Figure S2.

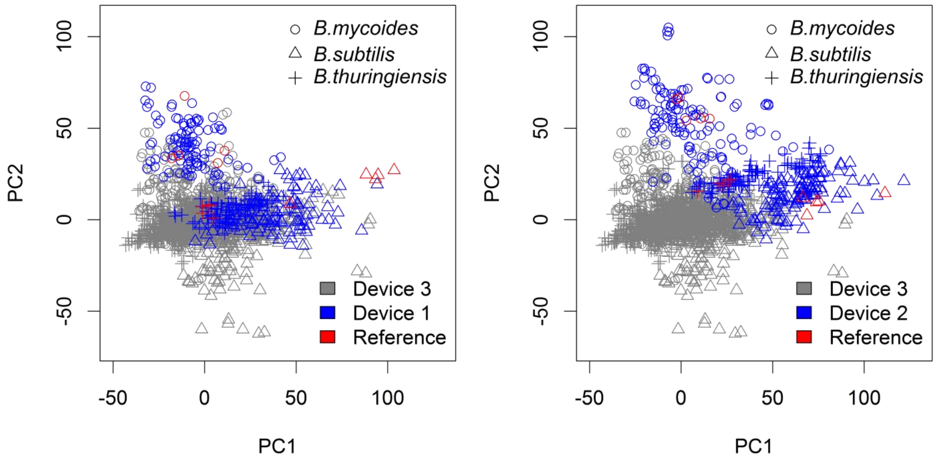


Figure S7 PC plots of the primary, the secondary, and the reference datasets. As shown in the plot on the left set, the reference samples of B. subtilis were strongly different as the center of the whole secondary dataset of the same group. This leads to unsatisfactory model transfer.

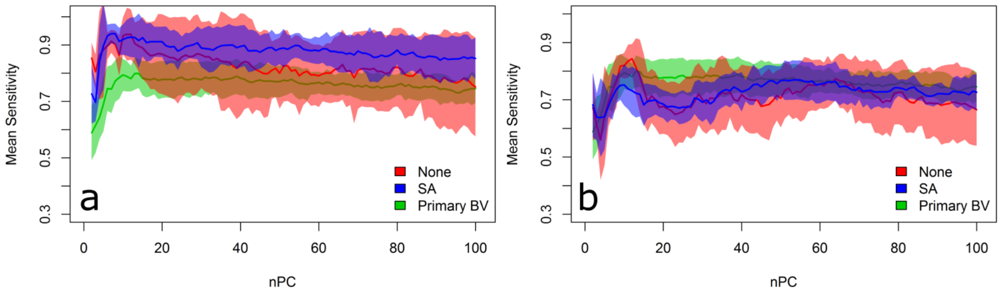


Figure S8 Similar results as shown in Figure 2d. Here the results of bacterial Raman spectra measured on the first and second devices are plotted separately in a and b, respectively.

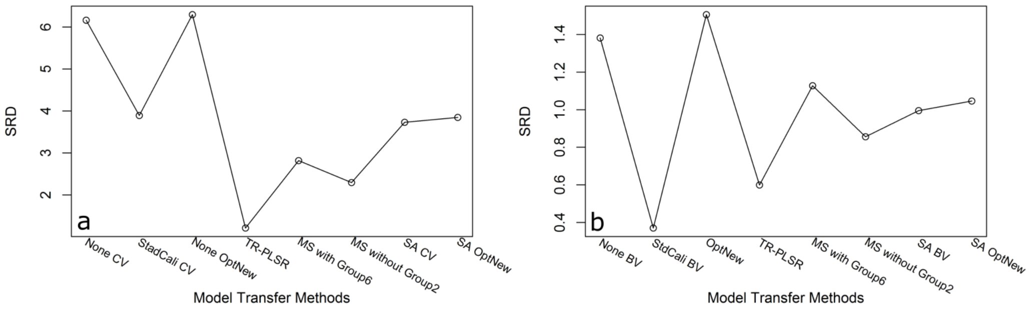


Figure S9. Results of SRDs: The SRD matrix was constructed with each column representing one model transfer method and each row corresponding to prediction from a different case of primary and secondary dataset. Finally, a 28×8 and 10×8 SRD matrix was obtained for spore and bacterial dataset, respectively. The SRD calculation was performed with the maxima of rows as target vector. **a**: Results from spore datasets. **b**: Results from bacterial datasets. The power of model transfer was shown in both cases for TR-PLSR, MS and SA methods. TR-PLSR gave superior prediction to MS and SA. In addition, MS without group achieved better model transfer than MS with group.

7.6 EMSC based Model Transfer for Raman Spectroscopy in Biological Applications (A6)

S. Guo, A. Kohler, B. Zimmermann, R. Heinke, S. Stöckel, P. Rösch, J. Popp, and T. Bocklitz, *Analytical Chemistry*, 2018, 90, 9787-9795.
(https://pubs.acs.org/doi/full/10.1021/acs.analchem.8b01536)

Der Nachdruck der folgenden Publikation erscheint mit freundlicher Genehmigung von Analytical Chemistry, Copyright [2018] American Chemical Society. Reproduced with permission from Analytical Chemistry, Copyright [2018] American Chemical Society.
(https://pubs.acs.org/doi/full/10.1021/acs.analchem.8b01536)

Erklärungen zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation

EMSC based model transfer for Raman spectroscopy in biological applications, S. Guo[1], A. Kohler[2], B. Zimmermann[3], R. Heinke[4], S. Stöckel[5], P. Rösch[6], J. Popp[7], and T. Bocklitz[8], *Analytical Chemistry*, 2018, 90, 9787-9795.

Beteiligt an (*Zutreffendes ankreuzen*)								
	1	2	3	4	5	6	7	8
Konzeption des Forschungsansatzes	x						x	x
Planung der Untersuchungen	x	x	x	x	x	x	x	x
Datenerhebung				x	x	x		
Datenanalyse und -interpretation	x	x	x					x
Schreiben des Manuskripts	x	x	x			x	x	x
Vorschlag Anrechnung Publikationsäquivalente	1.0							

Downloaded via THURINGER UNIV LANDESBIBLIOTHEK on November 15, 2018 at 14:45:46 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

analytical chemistry
Article
Cite This: *Anal. Chem.* 2018, 90, 9787−9795
pubs.acs.org/ac

Extended Multiplicative Signal Correction Based Model Transfer for Raman Spectroscopy in Biological Applications

Shuxia Guo,[†,‡] Achim Kohler,[§] Boris Zimmermann,[§] Ralf Heinke,[†] Stephan Stöckel,[†] Petra Rösch,[†] Jürgen Popp,[†,‡,∥] and Thomas Bocklitz*[,†,‡]

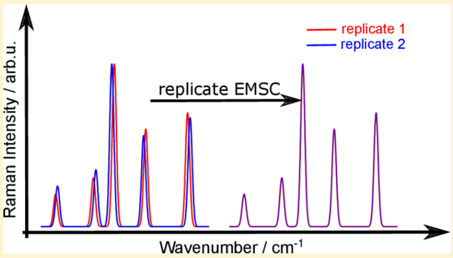
[†]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University of Jena, Helmholtzweg 4, D-07743 Jena, Germany

[‡]Leibniz Institute of Photonic Technology, Member of Leibniz Research Alliance 'Health Technologies', Albert-Einstein-Straße 9, D-07745 Jena, Germany

[§]Faculty of Science and Technology, Norwegian University of Life Sciences, P.O. Box 5003, NO1432, Ås, Norway

[∥]InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, D-07743 Jena, Germany

Supporting Information



ABSTRACT: The chemometric analysis of Raman spectra of biological materials is hampered by spectral variations due to the instrumental setup that overlay the subtle biological changes of interest. Thus, an established statistical model may fail when applied to Raman spectra of samples acquired with a different device. Therefore, model transfer strategies are essential. Herein we report a model transfer approach based on extended multiplicative signal correction (EMSC). As opposed to existing model transfer methods, the EMSC based approach does not require group information on the secondary data sets, thus no extra measurements are required. The proposed model-transfer approach is a preprocessing procedure and can be combined with any method for regression and classification. The performance of EMSC as a model transfer method was demonstrated with a data set of Raman spectra of three *Bacillus* bacteria spore species (*B. mycoides*, *B. subtilis*, and *B. thuringiensis*), which were acquired on four Raman spectrometers. A three-group classification by partial least-squares discriminant analysis (PLS-DA) with leave-one-device-out external cross-validation (LODCV) was performed. The mean sensitivities of the prediction on the independent device were considerably improved by the EMSC method. Besides the mean sensitivity, the model transferability was additionally benchmarked by the newly defined numeric markers: (1) relative Pearson's correlation coefficient and (2) relative Fisher's discriminant ratio. We show that these markers have led to consistent conclusions compared to the mean sensitivity of the classification. The advantage of our defined markers is that the evaluation is more effective and objective, because it is independent of the classification models.

Over the last decades, Raman spectroscopy has become one of the most versatile analytical techniques in biology and biomedicine, with applications in medical diagnostics,[1−8] microbe identification,[9−12] investigations of the metabolism,[13−15] and intraoperative decision making.[16] The fast development of Raman techniques is driven by several favorable properties of Raman spectroscopy.[17−20] First, Raman spectroscopy provides molecular fingerprint information on all biomolecules, making it ideally suitable for biological measurements. Second, Raman spectra can be easily obtained from aqueous solutions since water does not cause large interference, as opposed to infrared spectroscopy. Third, label-free and nondestructive measurements are possible with Raman spectroscopy, providing a great potential for in vivo investigations. Last but not the least, chemometric methods are able to distinguish the subtle spectral variations caused by biological changes, which is important to enhance the accuracy and speed of Raman based detections.

However, Raman-based biological applications can be hampered by measurement related variations, which make it difficult to establish robust models for classification. First, natural variations between the samples, such as patients or biological replicates, often create large spectral differences. This often hinders detection of biological effects of interest, such as effect of cancer, pathogens, or toxins on tissues and cells. Second, differences in measurement conditions often cause significant variations between the Raman spectra, even if they are from the same sample. Typical factors that cause variability are sample preparation and measurement protocols,

Received: April 5, 2018
Accepted: July 17, 2018
Published: July 17, 2018

ACS Publications © 2018 American Chemical Society

DOI: 10.1021/acs.analchem.8b01536
Anal. Chem. 2018, 90, 9787−9795

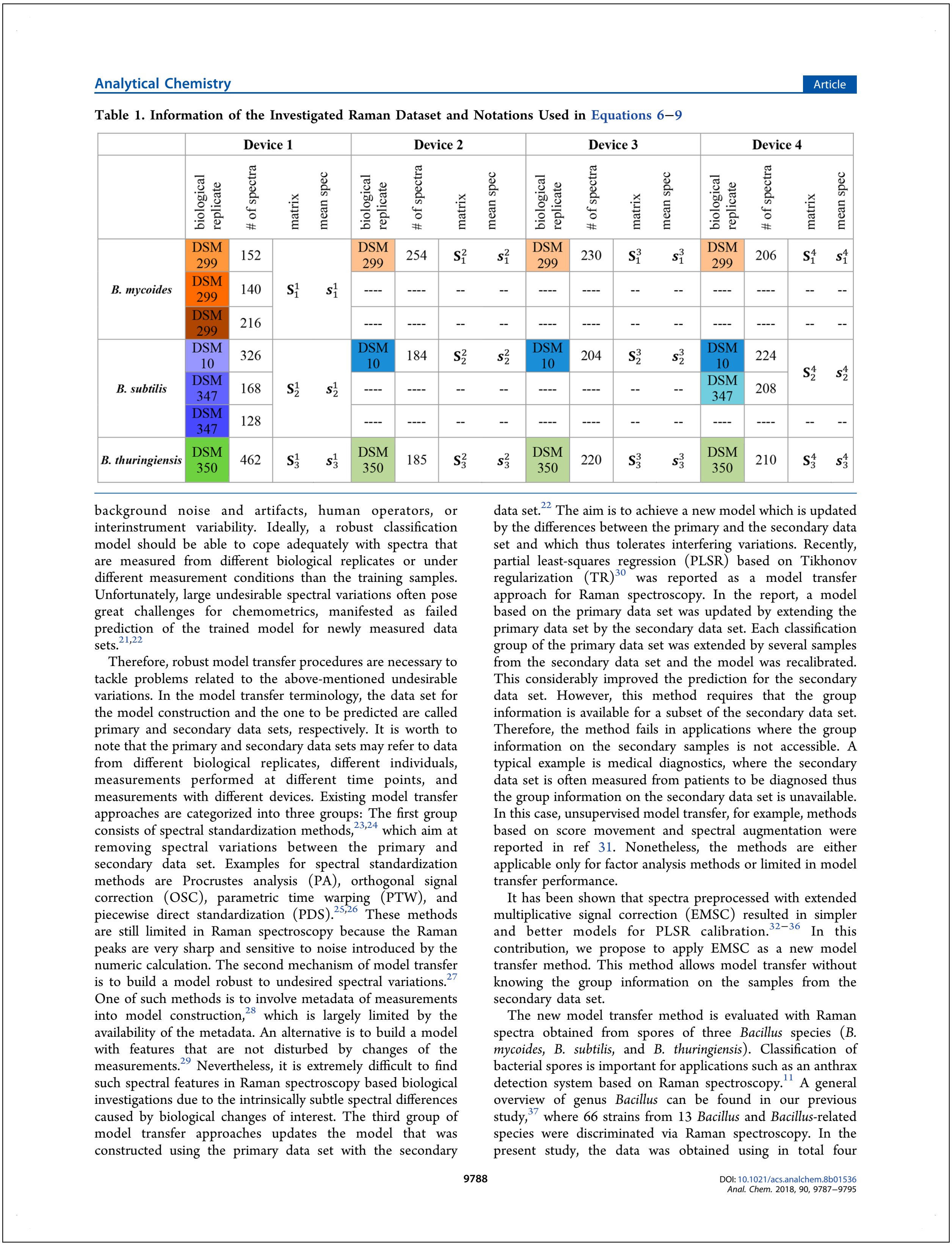


Table 1. Information of the Investigated Raman Dataset and Notations Used in Equations 6−9

	Device 1				Device 2				Device 3				Device 4			
	biological replicate	# of spectra	matrix	mean spec	biological replicate	# of spectra	matrix	mean spec	biological replicate	# of spectra	matrix	mean spec	biological replicate	# of spectra	matrix	mean spec
B. mycoides	DSM 299	152	$\mathbf{S}_1^1$	$\boldsymbol{s}_1^1$	DSM 299	254	$\mathbf{S}_1^2$	$\boldsymbol{s}_1^2$	DSM 299	230	$\mathbf{S}_1^3$	$\boldsymbol{s}_1^3$	DSM 299	206	$\mathbf{S}_1^4$	$\boldsymbol{s}_1^4$
	DSM 299	140			----	----	--	--	----	----	--	--	----	----	--	--
	DSM 299	216			----	----	--	--	----	----	--	--	----	----	--	--
B. subtilis	DSM 10	326	$\mathbf{S}_2^1$	$\boldsymbol{s}_2^1$	DSM 10	184	$\mathbf{S}_2^2$	$\boldsymbol{s}_2^2$	DSM 10	204	$\mathbf{S}_2^3$	$\boldsymbol{s}_2^3$	DSM 10	224	$\mathbf{S}_2^4$	$\boldsymbol{s}_2^4$
	DSM 347	168			----	----	--	--	----	----	--	--	DSM 347	208		
	DSM 347	128			----	----	--	--	----	----	--	--	----	----	--	--
B. thuringiensis	DSM 350	462	$\mathbf{S}_3^1$	$\boldsymbol{s}_3^1$	DSM 350	185	$\mathbf{S}_3^2$	$\boldsymbol{s}_3^2$	DSM 350	220	$\mathbf{S}_3^3$	$\boldsymbol{s}_3^3$	DSM 350	210	$\mathbf{S}_3^4$	$\boldsymbol{s}_3^4$

background noise and artifacts, human operators, or interinstrument variability. Ideally, a robust classification model should be able to cope adequately with spectra that are measured from different biological replicates or under different measurement conditions than the training samples. Unfortunately, large undesirable spectral variations often pose great challenges for chemometrics, manifested as failed prediction of the trained model for newly measured data sets.[21,22]

Therefore, robust model transfer procedures are necessary to tackle problems related to the above-mentioned undesirable variations. In the model transfer terminology, the data set for the model construction and the one to be predicted are called primary and secondary data sets, respectively. It is worth to note that the primary and secondary data sets may refer to data from different biological replicates, different individuals, measurements performed at different time points, and measurements with different devices. Existing model transfer approaches are categorized into three groups: The first group consists of spectral standardization methods,[23,24] which aim at removing spectral variations between the primary and secondary data set. Examples for spectral standardization methods are Procrustes analysis (PA), orthogonal signal correction (OSC), parametric time warping (PTW), and piecewise direct standardization (PDS).[25,26] These methods are still limited in Raman spectroscopy because the Raman peaks are very sharp and sensitive to noise introduced by the numeric calculation. The second mechanism of model transfer is to build a model robust to undesired spectral variations.[27] One of such methods is to involve metadata of measurements into model construction,[28] which is largely limited by the availability of the metadata. An alternative is to build a model with features that are not disturbed by changes of the measurements.[29] Nevertheless, it is extremely difficult to find such spectral features in Raman spectroscopy based biological investigations due to the intrinsically subtle spectral differences caused by biological changes of interest. The third group of model transfer approaches updates the model that was constructed using the primary data set with the secondary data set.[22] The aim is to achieve a new model which is updated by the differences between the primary and the secondary data set and which thus tolerates interfering variations. Recently, partial least-squares regression (PLSR) based on Tikhonov regularization (TR)[30] was reported as a model transfer approach for Raman spectroscopy. In the report, a model based on the primary data set was updated by extending the primary data set by the secondary data set. Each classification group of the primary data set was extended by several samples from the secondary data set and the model was recalibrated. This considerably improved the prediction for the secondary data set. However, this method requires that the group information is available for a subset of the secondary data set. Therefore, the method fails in applications where the group information on the secondary samples is not accessible. A typical example is medical diagnostics, where the secondary data set is often measured from patients to be diagnosed thus the group information on the secondary data set is unavailable. In this case, unsupervised model transfer, for example, methods based on score movement and spectral augmentation were reported in ref 31. Nonetheless, the methods are either applicable only for factor analysis methods or limited in model transfer performance.

It has been shown that spectra preprocessed with extended multiplicative signal correction (EMSC) resulted in simpler and better models for PLSR calibration.[32−36] In this contribution, we propose to apply EMSC as a new model transfer method. This method allows model transfer without knowing the group information on the samples from the secondary data set.

The new model transfer method is evaluated with Raman spectra obtained from spores of three *Bacillus* species (*B. mycoides*, *B. subtilis*, and *B. thuringiensis*). Classification of bacterial spores is important for applications such as an anthrax detection system based on Raman spectroscopy.[11] A general overview of genus *Bacillus* can be found in our previous study,[37] where 66 strains from 13 *Bacillus* and *Bacillus*-related species were discriminated via Raman spectroscopy. In the present study, the data was obtained using in total four different Raman spectrometers. A classification model was established by partial least-squares discriminant analysis (PLS-DA) that classified the samples into three groups. Validation was performed by leave-one-device-out cross-validation (LODCV). The model transferability was benchmarked by calculating the mean sensitivity of the prediction for each independent device. In addition, numeric markers were defined, which have the potential to be new benchmarks of the model transferability.

MATERIALS AND METHODS

Spores Cultivation and Raman Spectroscopy. The details on the cultivation and measurement of spores of *B. mycoides*, *B. thuringienses*, and *B. subtilis* can be found in our previous publication.[22] In summary, measurements were done on four micro Raman devices, which differed in their CCD detectors. Quartz was used as substrate for the samples measured on the first device, while nickel foil was utilized as substrate for the other three devices. The sample sizes of different species were relatively similar. Details are given in Table 1, where information on biological replicates were represented with different colors. To make the description clearer, we will refer to "biological replicate" as "batch" henceforth. Three batches of *B. mycoides* and *B. subtilis*, as well as one batch of *B. thuringienses* were measured with the first device, while only one batch per species was measured with the other three devices. In addition, the measurement with the latter three devices was done on the identical batches. An additional batch of *B. subtilis* (DSM 347) was measured on the fourth device.

This data set is suitable to investigate model transfer problems because of the following reasons. The three species feature very similar Raman spectra. The interspecies spectral differences are very subtle and smaller than interdevice differences (see Figure S1). Hence, a model transfer between different measurements is necessary. Meanwhile, the biological changes of the samples are negligible during the measurements because of the high stability and tolerance of the spores. This makes it possible to verify model transfer methods without being influenced by changes of the samples.

Spectral Analysis. The data analysis started with the spectral pretreatment. All Raman spectra were despiked with in-house written algorithms. The wavenumber axis was calibrated with the method described in ref 38 with 4-acetamidophenol as the standard material. Thereafter the baseline was corrected with two approaches, including the automatic optimization pipeline based on the previously defined marker R^{12}[39] (denoted as R^{12} optimization henceforth) and the basic EMSC.[33] Thereafter, an optional replicate correction was performed by the replicate EMSC, which was used as a model transfer procedure as described in the following section. As the last step of the preprocessing pipeline, all Raman spectra were vector normalized within the regions from 600 to 1750 cm^{-1} and 2800 to 3150 cm^{-1}. These regions were afterward used to construct a partial least-squares discriminant analysis (PLS-DA) model to classify the three spore species. A leave-one-device-out cross-validation (LODCV) was utilized as an external cross-validation, while the optimal number of latent variables (*nLV*) was optimized by a 5-fold internal cross-validation based on the training data set. The mean sensitivity of the prediction for each independent device was investigated as a benchmark of the model transferability. All computations were accomplished in statistical programming language Gnu R.[40]

Extended Multiplicative Signal Correction. The multiplicative signal correction (MSC)[33] model is given by

$$I(\tilde{\nu}) = a + b\cdot m(\tilde{\nu}) + e(\tilde{\nu}) \tag{1}$$

where a measured spectrum $I(\tilde{\nu})$ is modeled around a reference spectrum $m(\tilde{\nu})$. The parameter a designs a baseline effect and the unmodeled, mainly chemical effects, are contained in the residual $e(\tilde{\nu})$. As a reference spectrum, a mean or a typical scatter-free standard spectrum can be used. The parameters a and b are computed by a least-squares fitting procedure, where nonuniform weights may be used for different wavenumbers $\tilde{\nu}$ or spectral regions. A baseline-corrected and scaled spectrum is obtained according to

$$I_c(\tilde{\nu}) = (I(\tilde{\nu}) - a - e(\tilde{\nu}))/b \tag{2}$$

where $I_c(\tilde{\nu})$ is the corrected spectrum. The division by b aims to correct multiplicative scattering effects. In eq 2, this step may be omitted and replaced by a normalization that follows the MSC correction procedure. In the parameter estimation in eq 1, the reference spectrum is crucial and cannot be omitted. The MSC was applied for calibration transfer of NIR spectroscopy in ref 25 and was proven to perform comparably to PDS and OSC. A more advanced version of MSC is extended multiplicative signal correction (EMSC),[41] in which the MSC model is extended by additional terms such as polynomials or principal components. The addition of polynomials allows the correction of nonconstant baselines. When adding polynomial terms $\tilde{\nu}^1$, $\tilde{\nu}^2$, ... $\tilde{\nu}^n$, the EMSC model is written as

$$I(\tilde{\nu}) = a + b\cdot m(\tilde{\nu}) + d_1\tilde{\nu} + d_2\tilde{\nu}^2 + ... + d_n\tilde{\nu}^n + e(\tilde{\nu}) \tag{3}$$

where the parameters $d_1...d_n$ may be estimated by least-squares as for MSC. The spectra are thereafter corrected according to the same lines as shown in eq 2. The basic EMSC model refers to the case, where polynomials up to quadratic order are utilized for modeling in eq 3.

$$I(\tilde{\nu}) = a + b\cdot m(\tilde{\nu}) + d_1\tilde{\nu} + d_2\tilde{\nu}^2 + ... + d_n\tilde{\nu}^n + \sum_{k=1}^{N} g_k\cdot p_k(\tilde{\nu}) + e(\tilde{\nu}) \tag{4}$$

The EMSC model features a potential of further extension to remove other undesirable interferences. One of these extensions is the so-called replicate EMSC, which aims to decrease the inter-replicate spectral variations. As given in eq 4, the inter-replicate spectral variations are represented with the additional loading vectors $p_k(\tilde{\nu})$ obtained from, for example, a principal component analysis (PCA) on data from multiple replicates. After removing inter-replicate spectral variations, a model trained on certain replicates can be used to predict new replicates successfully. Herein "replicate" may refer to different batches (biological replicates), different individuals, or measurements from different devices. In this study, we term measurements from different devices "replicate" and ignore the batch information (see Table 1), unless otherwise stated. The loadings $p_k(\tilde{\nu})$ were computed according to two schemes in our investigation, as shown in Figure 1a,b.

(a) For the scheme of Figure 1a, data obtained from the same spore type with the same device were treated as one

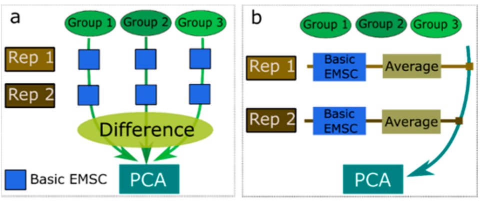


Figure 1. Graphic workflow of the two schemes to calculate loadings used in replicate EMSC model. (a) Replicate EMSC with group information: data measured from the same group and with the same device are used as one replicate. Each replicate is corrected with a basic EMSC using the mean spectrum of this replicate as the reference. Afterward the mean spectrum is calculated for each replicate and the difference between the mean spectra is computed for each two replicates belonging to the same group. These difference spectra are combined into a matrix and a PCA model was performed on this matrix. The resulted loadings are used as $p_k(\tilde{\nu})$ for the replicate EMSC model. (b) Replicate EMSC without group information: data measured on the same device are treated as one single replicate ignoring the group information. A basic EMSC is performed on each replicate using the mean spectrum of this replicate as the reference. Thereafter, the mean spectrum of each replicate is collected into a matrix and a PCA model is built with this matrix after column-wise mean-centering. The resulting loadings are used as $p_k(\tilde{\nu})$ in the replicate EMSC model.

replicate. A basic EMSC correction is performed for each replicate using the average of this replicate as the reference spectrum in eq 3. The mean spectrum of the resulting spectra is then calculated for each replicate. Thereafter the difference spectrum was calculated between every two mean spectra belonging to the same spore type but measured on different devices. At the end, all difference spectra were combined and a PCA was performed.

(b) For the scheme of Figure 1b, data measured on the same device were considered as one single replicate, regardless of which spore type they belonged to. Similar as for Figure 1a, the mean spectrum of each replicate was calculated after conducting a basic EMSC. Afterward all the mean spectra were combined and a PCA was performed after a column-wise mean-centering.

In both schemes of Figure 1a,b, the resulting loadings of the first N components are used as loading vectors ($p_k(\tilde{\nu})$ in eq 4 and a replicate EMSC model is constructed with all the spectra not corrected by basic EMSC. After estimation of the parameters in eq 4 by least-squares, the spectra are corrected according to

$$I_c(\tilde{\nu}) = \left(I(\tilde{\nu}) - a - d_1\tilde{\nu} - d_2\tilde{\nu}^2 - \ldots - d_n\tilde{\nu}^n - \sum_{k=1}^{N} g_k \cdot p_k(\tilde{\nu}) \right) / b \qquad (5)$$

The reason that both schemes in Figure 1a,b were investigated is the following. In principle, different spores of the same type contain almost identical biological characteristics if they were cultivated under the same growth conditions. Therefore, the inter-replicate spectral variations are mainly caused by the measurement conditions, device change in our case. On this basis, the inter-replicate spectral variations were calculated for each spore type separately in Figure 1a. However, the scheme of Figure 1a is limited because it requires knowing the group information of data from all replicates. Hence it becomes incapable for model transfer problems of biological diagnostics, where the group information on a new patient (replicate) is unknown and should be predicted. Such limitation was tackled by the scheme of Figure 1b, which ignores the group information when conducting replicate EMSC. This is feasible in most biological investigations because the spectral variations caused by biological changes of interest are very subtle and typically smaller than inter-replicate deviations. This fact was indicated in Figure S1, where the inter-replicate spectral variations dominated the interspecies spectral differences according to the Pearson correlation coefficients. Therefore, the components $p_k(\tilde{\nu})$ calculated according to Figure 1b are dominated by the inter-replicate deviations. As a consequence, the inter-replicate variations are removed via replicate EMSC without significantly losing biological related information. Figure 1b is inapplicable if inter-replicate spectral changes are smaller than the interspecies spectral variations. This can be proven by the three batches (biological replicates) measured on the first device, if we use each batch as one replicate (see Figure S2). In this case, however, model transfer becomes not necessary since adequate prediction is highly possible even without model transfer, as shown by the prediction from leave-one-batch-out cross validation given in Figure S3. Hereafter, we will denote the two schemes in Figure 1a,b as replicate EMSC with group information and replicate EMSC without group information, respectively. The results of the two schemes were compared and presented in the Results and Discussion section.

Definition of Numeric Markers. When a data set includes multiple replicates, it is of great importance to evaluate the influence of the inter-replicate variations on the prediction results. This information can be used to judge if the inter-replicate variations are acceptable for a given classification task or if a model transfer is necessary. The most straightforward way to perform this evaluation is to construct a classification model with certain replicates and predict independent replicates. A successful prediction demonstrates that the inter-replicate variation is low compared to the intergroup variation and the model transfer is not necessary. This approach requires the construction of a classification model and therefore, the conclusion is largely dependent on the performance of the utilized classification model. To overcome this issue, we defined two markers to evaluate the inter-replicate variations with respect to intergroup differences. The definitions are based on Pearson's correlation coefficient and Fisher's discriminant ratio.[42,43]

Before the markers are introduced, it is necessary to clarify the notations used in the definitions. Table 1 summarizes the information on the data sets used in this investigation. Ignoring the information on batches, we denoted the spectra from group m (m = 1, 2, 3) measured on the ith (i = 1, 2, 3, 4) device (i.e., ith replicate) by a matrix $\mathbf{S}_m^i$, with each spectrum as a row. The corresponding mean spectrum is termed s_m^i. On this basis, the definitions of the markers are the following.

(1) Relative Pearson's correlation coefficient

Pearson's correlation coefficient between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma_x \sigma_y} + 1 \right) / 2 \qquad (6)$$

where cov($\boldsymbol{x}$, $\boldsymbol{y}$) denotes covariance between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, while σ_x and σ_y represents the standard deviation of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Accordingly, the relative Pearson's correlation coefficient (ρ_{mn}^{i}) between group m ($m = 1, 2, 3$) and group n ($n = 1, 2, 3, n \neq m$) from the ith ($i = 1, 2, 3, 4$) replicate is defined as

$$\rho_{mn}^{i} = \frac{\rho(\boldsymbol{s}_m^i, \boldsymbol{s}_n^i)}{\sqrt{\rho(\boldsymbol{s}_m^i, \boldsymbol{s}_m^{j(j\neq i)}) \cdot \rho(\boldsymbol{s}_n^i, \boldsymbol{s}_n^{j(j\neq i)})}}$$
$$(m, n = 1, 2, 3, m \neq n; i = 1, 2, 3, 4) \tag{7}$$

where $\boldsymbol{s}_m^{j(j\neq i)}$ and $\boldsymbol{s}_n^{j(j\neq i)}$ represents the mean spectrum of group m and n from all replicates except the ith replicate, respectively. In this way, we normalized the similarity between group m and group n for the ith replicate ($\rho(\boldsymbol{s}_m^i, \boldsymbol{s}_n^i)$) by the inter-replicate similarity of the two groups ($\rho(\boldsymbol{s}_m^i, \boldsymbol{s}_m^{j(j\neq i)})$ and $\rho(\boldsymbol{s}_n^i, \boldsymbol{s}_n^{j(j\neq i)})$).

(2) Relative Fisher's discriminant ratio

Fisher's discriminant ratio between matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is calculated by

$$d(\mathbf{X}, \mathbf{Y}) = \frac{2 \times \| \bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} \|_2}{\sum_{k=1}^{N_X} \| \boldsymbol{X}_k - \bar{\boldsymbol{X}} \|_2 / N_X + \sum_{k=1}^{N_Y} \| \boldsymbol{Y}_k - \bar{\boldsymbol{Y}} \|_2 / N_Y} \tag{8}$$

where $\bar{\mathbf{X}}$, $\boldsymbol{X}_k$, and N_X denotes the column-mean (mean spectrum), the kth row (spectrum), and the number of rows (spectra) related to matrix $\mathbf{X}$, respectively. The denotations are the same for $\bar{\mathbf{Y}}$, $\boldsymbol{Y}_k$, and N_Y in terms of matrix $\mathbf{Y}$. The relative Fisher's discriminant ratio (d_{mn}^{i}) between group m ($m = 1, 2, 3$) and group n ($n = 1, 2, 3, n \neq m$) from ith ($i = 1, 2, 3, 4$) replicate is defined as

$$d_{mn}^{i} = \frac{d(\mathbf{S}_m^i, \mathbf{S}_n^i)}{\sqrt{d(\mathbf{S}_m^i, \mathbf{S}_m^{j(j\neq i)}) \cdot d(\mathbf{S}_n^i, \mathbf{S}_n^{j(j\neq i)})}}$$
$$(m, n = 1, 2, 3, m \neq n; i = 1, 2, 3, 4) \tag{9}$$

where $\mathbf{S}_m^{j(j\neq i)}$ and $\mathbf{S}_n^{j(j\neq i)}$ gives the spectral matrix of group m and group n from all replicates except the ith replicate. Apparently, the d_{mn}^{i} is derived by normalizing the Fisher's discriminant ratio between group m and group n of the ith replicate ($d(\mathbf{S}_m^i, \mathbf{S}_n^i)$) by the inter-replicate Fisher's discriminant ratio discriminant of the two groups ($d(\boldsymbol{S}_m^i, \boldsymbol{S}_m^{j(j\neq i)})$ and $d(\boldsymbol{S}_n^i, \boldsymbol{S}_n^{j(j\neq i)})$).

Methods Validation. The proposed approaches were validated based on the Raman spectra of bacterial spores of three species measured with four devices. The EMSC was performed with the mean spectrum of the involved data as the reference spectrum. The mean intensity of the reference spectrum within the region 1800−2800 cm^{-1} was subtracted from the reference spectrum before EMSC modeling,[32] since this region does not show any significant Raman peaks of biological samples. The corrected spectra were used for a three-group classification by PLS-DA with LODCV. The prediction was benchmarked by mean sensitivity, which is defined as the average of the sensitivities of the three spore types (see Table S1).

As the first verification, the performance of model transfer was compared for different replicate EMSC mechanisms shown in Figure 1a,b. The impact of the replicate EMSC without group information on the spectra was investigated according to the interdevice and intergroup Pearson's correlation coefficients. A paired Mann−Whitney test was performed in both cases to compare the results before and after replicate EMSC.

Afterward, the performance of replicate EMSC as a model transfer tool was evaluated and compared to other preprocessing methods as well as the TR-based model transfer. The employed preprocessing approaches included R^{12} optimization,[39] basic EMSC, replicate EMSC,[33] and R^{12} optimization combined with replicate correction. The prediction from these methods was compared according to p values of a paired Mann−Whitney test and the values of sum of ranking difference (SRD).[44] The paired Mann−Whitney test was conducted between R^{12} optimization (i.e., without model transfer) and the other methods based on the 12 sensitivities (three sensitivities for each device). The null hypothesis was that the results of the R^{12} optimization were not less than the model transfer methods. The major pitfall of the hypothesis test was that the p value of a hypothesis test is strongly influenced by the sample size.[45] Given the 12 sensitivities in our investigation, the p value would be above 0.05, even if one single value satisfies the null hypothesis. To prove this statement, we split the Raman spectra of each device into five folds by a statistical resampling and calculated the sensitivities for each fold separately based on the prediction from LODCV. This resulted in 60 sensitivities (15 × 4), which were used for an additional Mann−Whitney test (namely, resampled test). In comparison to hypothesis test, SRD was less influenced by sample size and was proven to provide unambiguous results for comparing different methods.[44] The values were calculated on the basis of the 12 sensitivities for each method. A lower SRD demonstrated a better performance of the method.

Additional to the previous evaluation, replicate EMSC was also compared to PDS. PDS has been successfully applied for calibration transfer in NIR spectroscopy but not applied to Raman spectroscopic data. It works by standardizing primary and secondary spectra using piecewise principal component regression. In this contribution, the number of principal components was optimized automatically while the window size was manually selected to ensure an optimal standardization. Details of the computations are provided in the Supporting Information.

Another important validation is to compare the prediction after model transfer with the prediction by the model built on the secondary data set itself. To do so, we used the Raman spectra from the first device as the secondary data set, because it contained multiple batches (see Table 1). The classification was performed with leave-one-batch-out cross validation. The resulting sensitivities of the three species were compared to those from the prediction with the model built on Raman spectra from the other three devices.

Stability Test of Replicate EMSC. According to eqs 4 and 5, it is clear that the performance of the replicate EMSC is dependent on the reliability of $p_k(\tilde{\nu})$. Therefore, it is necessary to test the stability of the replicate EMSC, if $p_k(\tilde{\nu})$ is calculated from varying number of secondary samples belonging to different groups. This is extremely important for real world application, because the secondary data set to be predicted can be composed of a single group, or very few samples. Herewith we denote the secondary samples used for calculating $p_k(\tilde{\nu})$ as transfer samples.

For verification, we utilized the Raman spectra of the second device as the primary data set. The Raman spectra from the first device were used as secondary data set. The first device

DOI: 10.1021/acs.analchem.8b01536
Anal. Chem. 2018, 90, 9787−9795

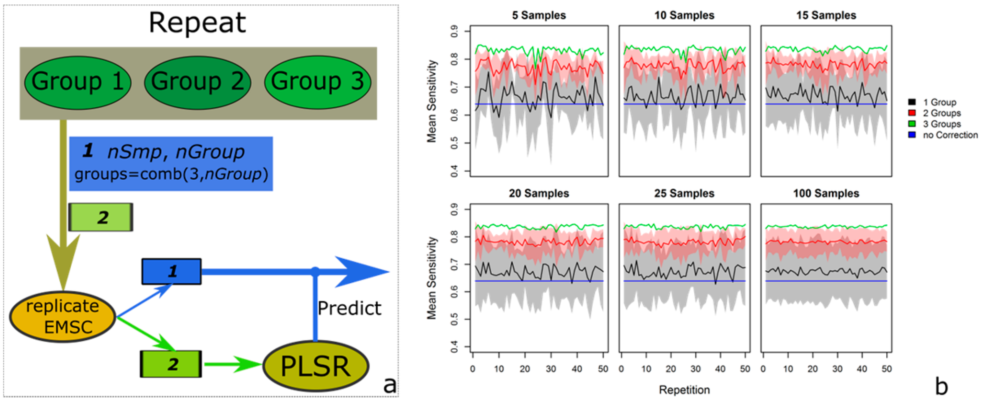


Figure 2. Graphic workflow and results of stability test. (a) Raman spectra from the second and the first devices were used as primary and secondary data set, respectively. Different numbers of transfer samples (*nSmp*) were selected from different group combinations. The transfer samples and all the primary samples were used to build a replicate EMSC model, which was utilized to correct the primary and the secondary data set. Thereafter, a PLSR model was trained with the corrected primary samples and utilized to predict all the secondary samples. These steps were repeated for 50 times giving each combination of *nSmp* and *nGroup*. (b) Mean sensitivities of the prediction on the secondary data set. Each subplot represents a different *nSmp*, where each series corresponded to the same *nGroup* with various group combinations. The average and the standard deviation of the results are visualized as solid lines and shades, respectively. The prediction of the secondary data set without replicate correction was provided for comparison. Accordingly, the replicate EMSC could increase the mean sensitivities regardless of the selected *nSmp* and *nGroup*. However, a strong improvement was observed, if transfer samples are composed of multiple groups. On the other hand, the value of *nSmp* did not have a strong influence on the prediction, given the same *nGroup*. Nevertheless, the performance was more stable over the repeated samplings for larger *nSmp*.

was utilized as secondary data set because the samples measured on the first device were composed of multiple batches (Table 1). This provided larger intragroup diversity than those from the other three devices. Thus, the repeated sampling for the transfer samples is more meaningful.

The workflow of the test is visualized in Figure 2a. Hereby, we randomly selected different number of transfer samples (*nSmp*) from the secondary data set composed of different number of groups (*nGroup*). There are seven possible group combinations: C_3^1 for $nGroup = 1$, C_3^2 for $nGroup = 2$, and C_3^3 for $nGroup = 3$. The value of *nSmp* was varied within [5, 10, 15, 20, 25, and 100]. We repeated the sampling 50 times for each combination of *nSmp* and *nGroup*. The $p_k(\tilde{\nu})$ were calculated with all the primary samples and the transfer samples. Then the coefficients of the replicate EMSC model were calculated with all primary samples and the transfer samples. The replicate correction was done for all primary and secondary data sets using the resulted replicate EMSC model. Thereafter a PLSR model was trained with the corrected primary data set and the corrected secondary data set was predicted. All these steps were repeated for 50 times giving each combination of *nSmp* and *nGroup*.

■ RESULTS AND DISCUSSION

The results from the aforementioned experiments are presented and discussed in this section.

Model Transferability. In this subsection, the performance of the replicate EMSC for model transfer was validated. First, the replicate EMSC was performed according to the two schemes shown in Figure 1, using the first loading of the PCA in both cases. The corrected Raman spectra were used for the three-group classification. The results of the prediction on the data of each device are visualized in Figure S4. Apparently, the results from the replicate EMSC without group information are comparable to the results obtained from the replicate EMSC with group information. The two schemes led to similar PCA loadings $p_k(\tilde{\nu})$. Additionally, the mean Raman spectra of each group measured on each device as well as the interspecies and interdevice Pearson's correlation coefficient were plotted in Figure S5, before and after replicate EMSC using the scheme of Figure 1b. The intensity was obviously changed after the replicate EMSC for the Raman spectra measured on the first device. This was because the substrate used in the measurements with this device was different from the substrates used for the measurements with the other three devices, leading to large intensity differences between Raman spectra of the first device to the others. After the replicate EMSC, these differences were removed, resulting in noticeable spectral changes. On the other hand, no significant spectral changes were caused by the replicate EMSC for the other three devices. Therefore, the scheme of Figure 1b, which does not need the group information, is a feasible alternative to Figure 1a, which requires the group information. The hypothesis test on the correlation coefficients (Figure S5) demonstrated that both interspecies and interdevice spectral differences were decreased after replicate EMSC. However, as is shown in subsection "values of the markers", the replicate EMSC is able to increase relative Pearson's correlation coefficient, i.e., the decrease in interdevice difference is larger than the decrease in interspecies difference. This ensures a successful model transfer with replicate EMSC. In the following, we will apply the replicate EMSC without group information, unless otherwise stated.

The mean sensitivities of the prediction resulted from replicate EMSC and other methods are summarized in Table 2. The first two columns show the results where the baseline was removed with the *R*[12] optimization and basic EMSC,

Table 2. Mean Sensitivity from Different Model Transfer Methods

	R^{12}	basic EMSC	TR	rep EMSC1	rep EMSC2	rep EMSC3	rep EMSC4	R^{12} EMSC1	R^{12} EMSC2	R^{12} EMSC3	R^{12} EMSC4
device 1	0.642	0.833	0.859	0.833	0.857	0.916	0.924	0.812	0.864	0.893	0.892
device 2	0.671	0.821	0.775	0.865	0.957	0.959	0.959	0.765	0.915	0.929	0.931
device 3	0.916	0.962	0.877	0.891	0.886	0.934	0.934	0.916	0.929	0.934	0.934
device 4	0.730	0.894	0.833	0.960	0.968	0.958	0.958	0.926	0.947	0.947	0.947
p value1		0.010	0.175	0.049	0.063	0.051	0.051	0.093	0.071	0.060	0.060
p value2		3.42×10^{-7}	2.95×10^{-3}	2.70×10^{-4}	2.08×10^{-4}	5.13×10^{-5}	5.13×10^{-5}	5.72×10^{-4}	8.70×10^{-5}	2.90×10^{-5}	3.04×10^{-5}
SRD	2.766	1.112	1.613	0.999	0.642	0.346	0.321	1.386	0.681	0.532	0.531

respectively. The third column gives the results of the TR, where the baseline was corrected with the R^{12} optimization. The fourth column presents the results from the replicate EMSC using the first loading of the PCA in the replicate EMSC model. The *p* values from the Mann−Whitney test without and with resampling were given by "*p* value1" and "*p* value2", respectively. The last row shows the results of SRD. The "*p* value1" were generally above 0.05, because of the small sample size.[45] This statement can be proven by the "*p* value2", where a resampling was conducted to enlarge the sample size. In this case, the *p* values were all below 0.0029, demonstrating a significant improvement of the prediction after model transfer. This experiment shows that the hypothesis test is not optimal to compare different model transfer methods in our case. The superiority of replicate EMSC was further evident according to SRD, which was the minimal for replicate EMSC method.

The results of the fourth column were obtained by only including one loading in the replicate EMSC model. The satisfying results demonstrated that a model transfer is possible even if only two replicates are measured, i.e., two different devices are used. In order to study the prediction as a function of the number of loadings included in the replicate EMSC model, we varied the number of loadings from 2 to 4 for the replicate EMSC model. Since our data was obtained from four devices, four components was the maximum number of components and corresponded to a full rank model. The results of PLS-DA with a LODCV are displayed in Table 2, labeled as "repEMSC *n*" (n = 2, 3, 4). It can be observed, that the prediction results improved when using more than one loading. Nonetheless, including too many loadings in the replicate EMSC model may lead to loss of biological related information and lead to reduced prediction ability.

In practice, the number of loadings for replicate EMSC model can be optimized by cross-model-validation, as visualized in Figure S6. Thereby one device is taken out as a secondary device. A replicate EMSC is performed based on both primary and secondary data sets with a given number of loadings (*nComp*). Afterward, the corrected primary data set is used to train a PLS-DA model, where a cross-validation (CV) is performed and the number of latent variables is optimized. The averaged validation mean sensitivity is recorded as a benchmark of *nComp*. The classification is repeated for each possible value of *nComp*. The *nComp* corresponding to the best averaged validation mean sensitivity is selected as optimal value. The calculations are repeated using each device once as secondary device.

Table 2 also shows that the basic EMSC featured better predictions compared to the baseline correction by R^{12} optimization. That means, the basic EMSC improved the model transferability. To check the efficiency of the replicate correction terms $\sum_{k=1}^{N} g_k \cdot p_k(\tilde{\nu})$ in model transfer, we replaced the implied basic EMSC in the replicate EMSC model with the R^{12} optimization. Again, we varied the number of loadings in the replicate EMSC model. The results of the LODCV were shown in Table 2 labeled as "R^{12} EMSC *n*" (n = 1, 2, 3, 4), which were only slightly inferior to the normal replicate EMSC. That means the pure replicate correction is highly efficient in model transfer. This provides the possibility to combine a pure replicate correction with other baseline correction methods superior to the basic EMSC.

In addition, the results of comparison between replicate EMSC and PDS were visualized in Figure S7 and Table S2. It was shown that the PDS led to obvious signal loss of Raman bands and inferior prediction of the classification model than the replicate-EMSC. No improvement was seen by changing the window size of the piecewise regression. This fact indicates that PDS works better for broad NIR-peaks than for sharp peaks in Raman spectra.

The results of the leave-one-batch-out cross-validation on the first device and the prediction with the model built on Raman spectra from the other three devices were shown in Figure S3, where the mean, maximum, and minimum of the three sensitivities were visualized for each case of prediction. Surprisingly, the replicate EMSC achieved even better prediction than the local model built on the first device, especially if multiple loadings were included in the replicate EMSC model.

Noteworthy, the reference spectrum in the EMSC models was calculated from the mean spectrum without baseline correction in the previous experiments. When additional baseline correction was performed on the reference spectrum, the model transferability was slightly improved as can be seen in Figure S8.

Results of Stability Test. The mean sensitivities calculated according to the workflow in Figure 2a were plotted in Figure 2b, where each subplot represents a different value of *nSmp*. Every subplot includes three series, each corresponding to different group combination with the same value of *nGroup*. Solid lines and the shades represent the average and the standard deviation of the results, respectively. Compared to the prediction without replicate EMSC, the prediction was significantly improved by replicate EMSC regardless of the value of *nSmp* and *nGroup*. However, the improvement was dramatically enhanced, if the transfer samples were composed of two or more groups. On the other hand, the results were almost independent of *nSmp*, given the same *nGroup*. Nevertheless, the stability over the repeated samplings was enhanced by larger *nSmp*.

Values of the Markers. In the sections above, it was shown that the Raman spectra that were preprocessed by different preprocessing methods led to different model transferability (see Table 2). In this section, the model transferability was additionally evaluated with the defined markers. The markers d^i_{mn} (relative Fisher's discriminant ratio) and ρ^i_{mn} (relative Pearson's correlation coefficient) were calculated based on the Raman spectra that were corrected by different preprocessing methods. The results are visualized in Figures S9 and S10. For the sake of clarity, classification results corresponding to different preprocessing methods were replotted in Figure S9f. As it can be observed, d^i_{mn} increased from Figure S9a−e, indicating an improvement of model transferability. The results of the marker ρ^i_{mn} led to a similar conclusion, where decreased values illustrated higher model transferability. It is obvious that the markers defined in this paper are able to judge the model transferability independent of the applied classification model. Therefore, the markers may be used as potential benchmarks for preprocessing procedures from a model transfer perspective. Noteworthy, it is possible to first perform a PCA on the data sets and calculate d^i_{mn} and ρ^i_{mn} from the score vectors. The results are shown in Figure S11, which lead to similar conclusions as Figures S9 and S10.

■ CONCLUSION

We reported about a model transfer approach based on replicate EMSC and verified its capability for Raman spectroscopy in biological applications. The model transferability was dramatically improved using the replicate EMSC approach. Applications of this EMSC based model transfer method to other spectroscopic data are possible and will be researched. Comparing to the Tikhonov regularization (TR) method, the new method is superior in three aspects. First, a better model transferability was achieved by the replicate EMSC according to the mean sensitivities of the studied classification model. Second, the replicate EMSC does not require label information on the samples measured on the secondary device. Third, the replicate EMSC is a preprocessing method and is independent of the afterward applied statistical analysis techniques. Moreover, we defined two numeric markers, namely, the relative Pearson correlation coefficient and the relative Fisher's discriminant ratio, which gave consistent evaluation of the model transferability compared to the mean sensitivities of the classification. This means that the influence of inter-replicate variations on classification systems can be estimated by these markers without training a classification model. Furthermore, the markers can be utilized in two aspects: (1) The markers can evaluate the goodness of a spectral preprocessing from the perspective of model transferability. (2) The influence of inter-replicate deviations on the classification tasks can be estimated. In both cases the developed markers are useful and should be applied in the design phase of an experiment.

In further studies we will elucidate the performance of the replicate-EMSC as a model transfer tool encountering large inter-replicate spectral variations. These issues are important if the measurement devices feature strongly differing optical characteristics such as numerical aperture, excitation wavelengths, and groove density of the grating.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b01536.

> Inter-species and inter-device differences benchmarked with Pearson correlation coefficients; inter-species and inter-batch differences benchmarked with Pearson correlation coefficients calculated on the data measured with the first device; comparison of the two schemes for calculating the loadings in the replicate EMSC model; calculation of mean sensitivity; results before and after replicate EMSC procedure; graphic workflow of cross-model-validation; average of the reference spectra, average of the query spectra, and average of spectra after PDS for each device; mean sensitivity for the prediction of Raman spectra measured on different devices; sensitivities of the three species in different cases of prediction for the Raman spectra measured on the first device; comparison of the mean sensitivities for the prediction of each independent device; results of relative Fisher's discrimination, with Raman spectra corrected by different pre-processing procedures; results of relative Pearson's correlation coefficient, with Raman spectra corrected by different preprocessing procedures; and results of the markers, with Raman spectra corrected by different pre-processing procedures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: thomas.bocklitz@uni-jena.de. Fax: +49-3641-948302. Phone: +49-3641-948328.

ORCID

Jürgen Popp: 0000-0003-4257-593X
Thomas Bocklitz: 0000-0003-2778-6624

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The scholarship for S.G. from the “China Scholarship Council” (CSC) is greatly appreciated. Funding from the EU and BMBF via the project “PhotoSkin” (Grant FKZ 13N13243), support of EU via projects “HemoSpec” (Grants FP 7, CN 611682), and funding of the BMBF via the projects “Intersept” (Grant FKZ 13N13852) and “Uro-MDD” (Grant FKZ 03ZZ0444J) are highly acknowledged. The grant (Grant COST-STSM-BM1401-35148) of the COST action Raman4Clinics for S.G. is highly appreciated.

■ REFERENCES

(1) Matousek, P.; Stone, N. *Chem. Soc. Rev.* **2016**, *45*, 1794−1802.
(2) Barroso, E.; Smits, R.; Bakker Schut, T.; Ten Hove, I.; Hardillo, J.; Wolvius, E.; Baatenburg de Jong, R.; Koljenovic, S.; Puppels, G. *Anal. Chem.* **2015**, *87*, 2419−2426.
(3) Meyer, T.; Bergner, N.; Bielecki, C.; Krafft, C.; Akimov, D.; Romeike, B. F.; Reichart, R.; Kalff, R.; Dietzek, B.; Popp, J. *J. Biomed. Opt.* **2011**, *16*, 021113.
(4) Santos, I. s. P.; Caspers, P. J.; Bakker Schut, T. C.; van Doorn, R.; Noordhoek Hegt, V.; Koljenović, S.; Puppels, G. J. *Anal. Chem.* **2016**, *88*, 7683−7688.
(5) Kong, K.; Kendall, C.; Stone, N.; Notingher, I. *Adv. Drug Delivery Rev.* **2015**, *89*, 121−134.
(6) Pallaoro, A.; Hoonejani, M. R.; Braun, G. B.; Meinhart, C. D.; Moskovits, M. *ACS Nano* **2015**, *9*, 4328−4336.
(7) Wang, W.; Zhao, J.; Short, M.; Zeng, H. *J. Biophotonics* **2015**, *8*, 527−545.
(8) Vogler, N.; Bocklitz, T.; Subhi Salah, F.; Schmidt, C.; Bräuer, R.; Cui, T.; Mireskandari, M.; Greten, F. R.; Schmitt, M.; Stallmach, A. *J. Biophotonics* **2016**, *9*, 533−541.
(9) Berry, D.; Mahfoudh, K. B.; Wagner, M.; Loy, A. *Appl. Environ. Microbiol.* **2011**, *77*, 7846−7849.
(10) Walter, A.; Kuhri, S.; Reinicke, M.; Bocklitz, T.; Schumacher, W.; Rösch, P.; Merten, D.; Büchel, G.; Kothe, E.; Popp, J. *J. Raman Spectrosc.* **2012**, *43*, 1058−1064.
(11) Stöckel, S.; Meisel, S.; Elschner, M.; Rösch, P.; Popp, J. *Angew. Chem., Int. Ed.* **2012**, *51*, 5339−5342.
(12) Lorenz, B.; Wichmann, C.; Stöckel, S.; Rösch, P.; Popp, J. *Trends Microbiol.* **2017**, *25*, 413−424.
(13) Kumar B. N., V.; Guo, S.; Bocklitz, T.; Rösch, P.; Popp, J. *Anal. Chem.* **2016**, *88*, 7574−7582.
(14) Berry, B.; Moretto, J.; Matthews, T.; Smelko, J.; Wiltberger, K. *Biotechnol. Prog.* **2015**, *31*, 566−577.
(15) El-Mashtoly, S. F.; Petersen, D.; Yosef, H. K.; Mosig, A.; Reinacher-Schick, A.; Kötting, C.; Gerwert, K. *Analyst* **2014**, *139*, 1155−1161.
(16) Leblond, F.; Jermyn, M.; Mok, K.; Mercier, J.; Desroches, J.; Pichette, J.; Saint-Arnaud, K.; Guiot, M.-C.; Petrecca, K. *Sci. Transl. Med.* **2015**, *7*, 274ra19.
(17) Bocklitz, T. W.; Guo, S.; Ryabchykov, O.; Vogler, N.; Popp, J. *Anal. Chem.* **2016**, *88*, 133−151.
(18) Talari, A. C. S.; Movasaghi, Z.; Rehman, S.; Rehman, I. u. *Appl. Spectrosc. Rev.* **2015**, *50*, 46−111.
(19) Butler, H. J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N. J.; Gardner, B.; Martin-Hirsch, P. L. *Nat. Protoc.* **2016**, *11*, 664−687.
(20) Popp, J. *Ex-Vivo and in-Vivo Optical Molecular Pathology*; Wiley-Blackwell: Weinheim, Germany, 2014.
(21) Shahbazikhah, P.; Kalivas, J. H. *Chemom. Intell. Lab. Syst.* **2013**, *120*, 142−153.
(22) Guo, S.; Heinke, R.; Stöckel, S.; Rösch, P.; Bocklitz, T.; Popp, J. *Vib. Spectrosc.* **2017**, *91*, 111−118.
(23) Liang, C.; Yuan, H.-f.; Zhao, Z.; Song, C.-f.; Wang, J.-j. *Chemom. Intell. Lab. Syst.* **2016**, *153*, 51−57.
(24) Bocklitz, T.; Dörfer, T.; Heinke, R.; Schmitt, M.; Popp, J. *Spectrochim. Acta, Part A* **2015**, *149*, 544−549.
(25) Sjöblom, J.; Svensson, O.; Josefson, M.; Kullberg, H.; Wold, S. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 229−244.
(26) Fernández Pierna, J.; Boix Sanfeliu, A.; Slowikowski, B.; von Holst, C.; Maute, O.; Han, L.; Amato, G.; de la Roza Delgado, B.; Perez Marin, D.; Lilley, G. *Biotechnol., Agron., Soc. Environ.* **2013**, *17*, 547−555.
(27) Fearn, T. *J. Near Infrared Spectrosc.* **2001**, *9*, 229−244.
(28) Kalivas, J. H.; Siano, G. G.; Andries, E.; Goicoechea, H. C. *Appl. Spectrosc.* **2009**, *63*, 800−809.
(29) Kalivas, J. H.; Brownfield, B.; Karki, B. J. *J. Chemom.* **2017**, *31*, e2873.
(30) Kalivas, J. H. *J. Chemom.* **2012**, *26*, 218−230.
(31) Guo, S.; Heinke, R.; Stöckel, S.; Rösch, P.; Popp, J.; Bocklitz, T. *J. Raman Spectrosc.* **2018**, *49*, 627−637.
(32) Afseth, N. K.; Kohler, A. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 92−99.
(33) Liland, K. H.; Kohler, A.; Afseth, N. K. *J. Raman Spectrosc.* **2016**, *47*, 643−650.
(34) Martens, H.; Bruun, S. W.; Adt, I.; Sockalingum, G. D.; Kohler, A. *J. Chemom.* **2006**, *20*, 402−417.
(35) Zimmerman, B.; Tafintseva, V.; Bağcıoğlu, M.; Høegh Berdahl, M.; Kohler, A. *Anal. Chem.* **2016**, *88*, 803−811.
(36) Zimmermann, B.; Kohler, A. *Appl. Spectrosc.* **2013**, *67*, 892−902.
(37) Stöckel, S.; Meisel, S.; Elschner, M.; Rosch, P.; Popp, J. *Anal. Chem.* **2012**, *84*, 9873−9880.
(38) McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons: Toronto, Canada, 2000; Vol. *157*.
(39) Guo, S.; Bocklitz, T.; Popp, J. *Analyst* **2016**, *141*, 2396−2404.
(40) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.
(41) Kohler, A.; Sule-Suso, J.; Sockalingum, G.; Tobin, M.; Bahrami, F.; Yang, Y.; Pijanka, J.; Dumas, P.; Cotte, M.; Van Pittius, D. *Appl. Spectrosc.* **2008**, *62*, 259−266.
(42) Fisher, R. A. *Annals of Eugenics* **1936**, *7*, 179−188.
(43) Kohler, A.; Böcker, U.; Warringer, J.; Blomberg, A.; Omholt, S.; Stark, E.; Martens, H. *Appl. Spectrosc.* **2009**, *63*, 296−305.
(44) Héberger, K. *TrAC, Trends Anal. Chem.* **2010**, *29*, 101−109.
(45) Nakagawa, S.; Cuthill, I. C. *Biological reviews* **2007**, *82*, 591−605.

DOI: 10.1021/acs.analchem.8b01536
Anal. Chem. 2018, 90, 9787−9795

Supporting Information

Extended Multiplicative Signal Correction Based Model Transfer for Raman Spectroscopy in Biological Applications

Shuxia Guo[a,b], Achim Kohler[c], Boris Zimmermann[c], Ralf Heinke[a], Stephan Stöckel[a], Petra Rösch[a], Jürgen Popp[a,b,d] and Thomas Bocklitz*[a,b]

[a] Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University of Jena, Helmholtzweg 4, D-07743 Jena, Germany
[b] Leibniz Institute of Photonic Technology, Leibniz Research Alliance ‚Health Technologies‘, Albert-Einstein-Straße 9, D-07745 Jena, Germany
[c] Faculty of Science and Technology, Norwegian University of Life Sciences, PO Box 5003, NO1432 Ås, Norway
[d] InfectoGnostics, Forschungscampus Jena, Philosophenweg 7, D-07743 Jena, Germany
* Email: thomas.bocklitz@uni-jena.de; Fax: +49-3641-948302; Tel.: +49-3641-948328

Table of contents- workflow and results of spectral standardization and classification

Figure S1 Inter-species and inter-device differences benchmarked with Pearson correlation coefficients.

Figure S2 Inter-species and inter-batch differences benchmarked with Pearson correlation coefficients calculated on the data measured with the first device.

Figure S3 Comparison of the two schemes for calculating the loadings in the replicate EMSC model.

Table S1 Calculation of mean sensitivity.

Figure S4 Results before and after replicate EMSC procedure.

Figure S5 Graphic workflow of cross-model-validation.

Figure S6 Average of the reference spectra, average of the query spectra, and average of spectra after PDS for each device.

Table S2 Mean sensitivity for the prediction of Raman spectra measured on different devices.

Figure S7 Sensitivities of the three species in different cases of prediction for the Raman spectra measured on the first device.

Figure S8 Comparison of the mean sensitivities for the prediction of each independent device.

Figure S9 Results of relative Fisher’s discrimination, with Raman spectra corrected by different pre-processing procedures.

Figure S10 Results of relative Pearson’s correlation coefficient, with Raman spectra corrected by different pre-processing procedures.

Figure S11 Results of the markers, with Raman spectra corrected by different pre-processing procedures.

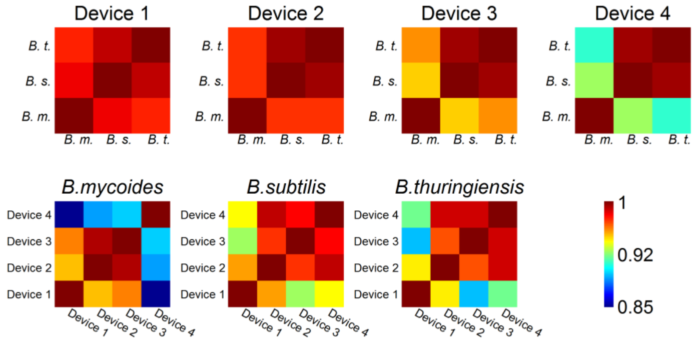


Fig. S1 Inter-species and inter-device differences benchmarked with Pearson correlation coefficients: The first row gives the results between the mean spectra of different species measured on the same device. The second row shows the results between the mean spectra of the same species measured on different devices. Apparently, the inter-device spectral variations are larger than inter-species variations.

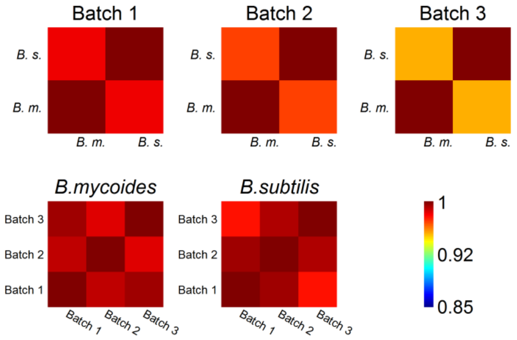


Fig. S2 Inter-species and inter-batch differences benchmarked with Pearson correlation coefficients calculated on the data measured with the first device: The first row gives the results between the mean spectra of different species belonging to the same batch. The second row shows the results between the mean spectra of the same species from different batches. The inter-batch spectral variations are smaller than the inter-species spectral variations for this dataset.

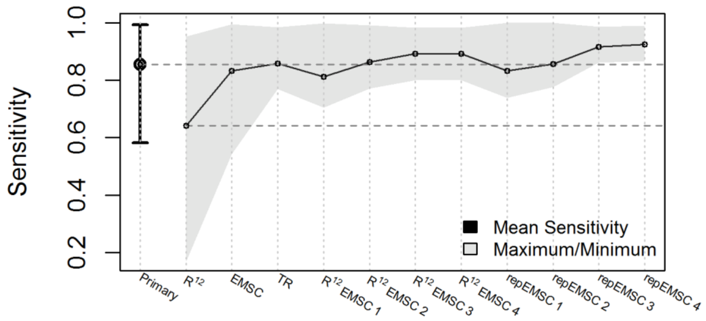


Fig. S3 Sensitivities of the three species in different cases of prediction for the Raman spectra measured on the first device, with the mean, maximum, and minimum visualized. The first column resulted from the prediction with model trained on the first device with a leave-one-batch-out cross-validation. The following columns are the prediction by the model built on Raman spectra of the other three devices with different model transfer methods.

Table S1 calculation of mean sensitivity. Term n_i^j means there are n samples of group i classified as group j

		Predicted		
		B. mycoides	*B. subtilis*	*B. thuringiensis*
True	*B. mycoides*	$\boldsymbol{n_1^1}$	n_1^2	n_1^3
	B. subtilis	n_2^1	$\boldsymbol{n_2^2}$	n_2^3
	B. thuringiensis	n_3^1	n_3^2	$\boldsymbol{n_3^3}$

$$\boldsymbol{sensitivity(i)} = \frac{\boldsymbol{n_i^i}}{\sum_{j=1}^{3} \boldsymbol{n_i^j}} \times \mathbf{100}\%$$

$$\boldsymbol{mean\ sensitivity} = \frac{\mathbf{1}}{\mathbf{3}} \cdot \sum_{i=1}^{3} \boldsymbol{sensitivity(i)}$$

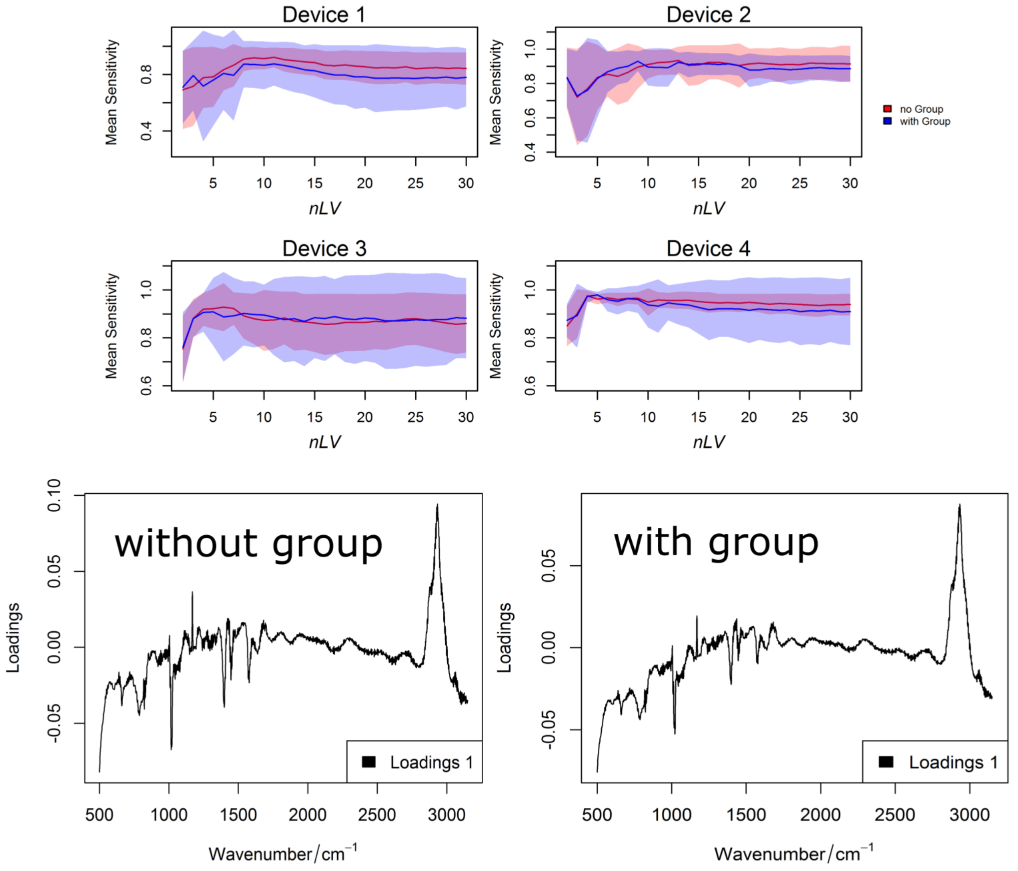
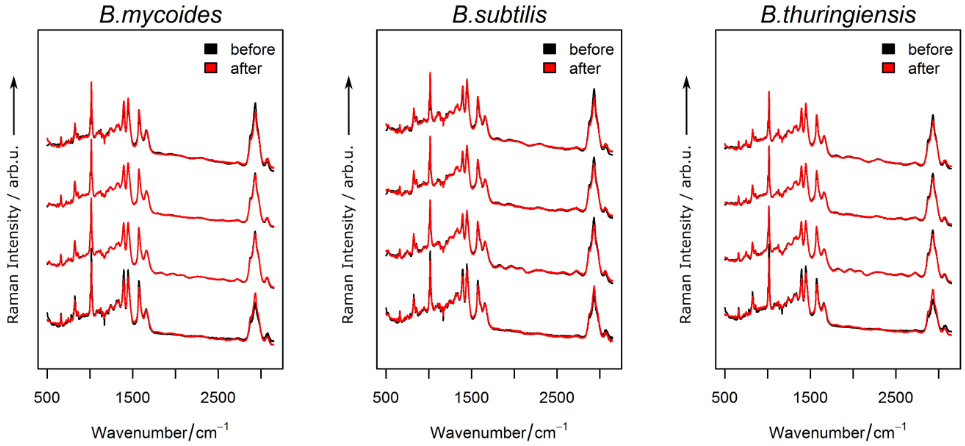
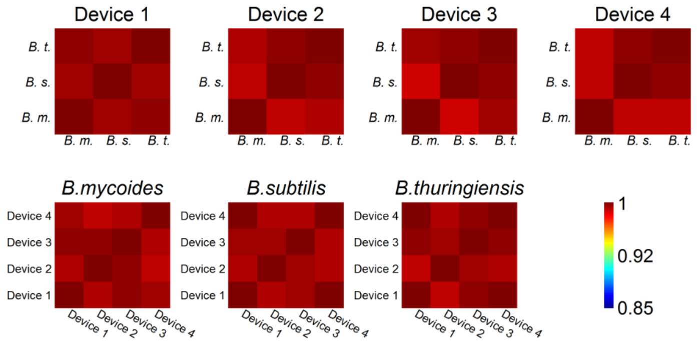


Fig. S4 Comparison of the two schemes for calculating the loadings in the replicate EMSC model, as shown in Fig. 1. The first loading was used for the replicate EMSC, which were quite similar for the two schemes. Meanwhile, the mean sensitivities of the prediction for each independent device were comparable; Mean sensitivity is depicted as a function of number of latent variables (*nLV*) in the PLS-DA model.



(a) Mean spectra of each group measured on each device, before and after replicate EMSC procedure.



(b) Inter-species and inter-device Pearson's correlation coefficients after replicate EMSC

Fig. S5 Results before and after replicate EMSC procedure. (a) The four spectra of each subplot are plotted with intensity offset to get clearer view. From bottom to up, the four spectra correspond to the four devices. No significant changes were observed for Raman spectra measured on the latter three devices. The changes for Raman spectra of the first device were much larger, because the samples and the substrate were different on this device as the others. Thus the original Raman spectra on the first device were greatly different to those of the other devices. (b) The first row gives the results between the mean spectra of different species measured on the same device. The second row shows the results between the mean spectra of the same species measured on different devices. The results were compared to Fig. S1 by a paired Mann-Whitney U test, with a null hypothesis of 'the correlation coefficient is decreased after replicate EMSC'. It was demonstrated that both inter-group and inter-device spectral variations are significantly decreased, with *p* values of 2.441e-4 and 3.815e-6, respectively. This is reasonable, since replicate EMSC works by making all spectra similar to the signal reference spectrum.

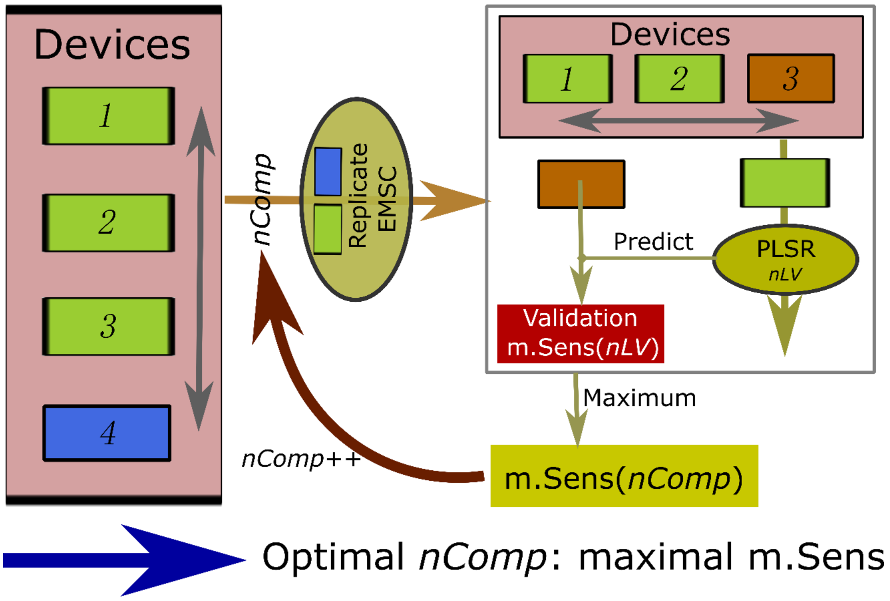


Fig. S6 Graphic workflow of cross-model-validation, where the number of loadings for replicate EMSC was optimized similar as the parameter of a classifier. Here the fourth device was used as the secondary device as an example. For each possible number of loadings (*nComp*), the replicate EMSC was performed on both the primary and the secondary datasets. Thereafter, a PLSR model was constructed on the primary dataset with a 5-fold cross-validation using different number of latent variables (*nLV*). The highest validation mean sensitivity was recorded. All these procedures were repeated for a new value of *nComp*. The optimal *nComp* was the one featuring the maximum average validation mean sensitivity. The procedures were the same if a new device was used as a secondary device.

Comparison of replicate EMSC to PDS and PTW

Piecewise direct standardization (PDS) removes inter-device spectral variations by transforming spectra measured by one device to those measured by the other device. This requires measuring a certain number of standard samples on both primary and secondary device. These spectra are referred to as reference spectra and are used to calculate the transformation matrix. Thereafter, the transformation matrix is applied on spectra of real samples to complete the spectral standardization. Because the transformation matrix is calculated with a moving window at local scale (piecewise), PDS can achieve better standardization than global approaches. Details of PDS can be found in *Chemom. Intell. Lab. Syst.* **32**, (1996) 201-213.

In our investigation, where more than two devices were involved, the reference spectra were calculated by averaging the spectra of all devices for each species separately. The standardization was conducted for each device individually, where the query spectra were composed of mean spectrum of each species measured by this device. The number of principal component for the piece-wise regression was optimized automatically so as to minimize the residual between the reference and corrected spectra. The width of the moving window was optimized by visual inspection.

Noteworthy, the Raman spectra from device 1 were excluded in this experiment because the substrate of the samples measured on this device are different as on the other devices. By using only the data from the other three devices, we could assume that differences between the reference and query spectra are merely from instrumental changes, which is an important assumption for PDS. The average of the reference spectra, average of the query spectra, and average of spectra after PDS were visualized for each device in Fig. S7. Obviously, the PDS led to severe signal loss of Raman bands for the fingerprint region (500-1800 cm^{-1}) and the C-H stretching region (2800-3150 cm^{-1}). This originates from two possible reasons: the spectral differences between the reference and query spectra are too large to be standardized reasonably; the Raman peaks are much sharper than near-infrared (NIR) spectroscopy and hampers the performance of PDS. For the silent region (1800-2800 cm^{-1}), where there are only broad Raman bands changing slowly, a good match is achieved between the query and corrected spectra.

After PDS, the corrected Raman spectra were used to build a three-group classifier, in which a leave-one-device-out cross-validation was applied to get the prediction. The mean sensitivities (see Table S2) were calculated to benchmark the prediction. The results were summarized in Table S2, in which the results without model transfer and with replicate EMSC were provided as a comparison. It is evident that the PDS resulted in a decrease of the prediction comparing to the results without PDS. The performance of replicate EMSC is obviously superior to PDS, regardless how many loadings ($p_k(\tilde{v})$) were included in the EMSC model.

Table S2 Mean sensitivity for the prediction of Raman spectra measured on different devices.

	R^{12}	**PDS**	**repEMSC1**	**repEMSC2**	**repEMSC3**
Device 2	0.881	0.807	0.835	0.961	0.956
Device 3	0.875	0.745	0.909	0.951	0.942
Device 4	0.810	0.776	0.961	0.926	0.926

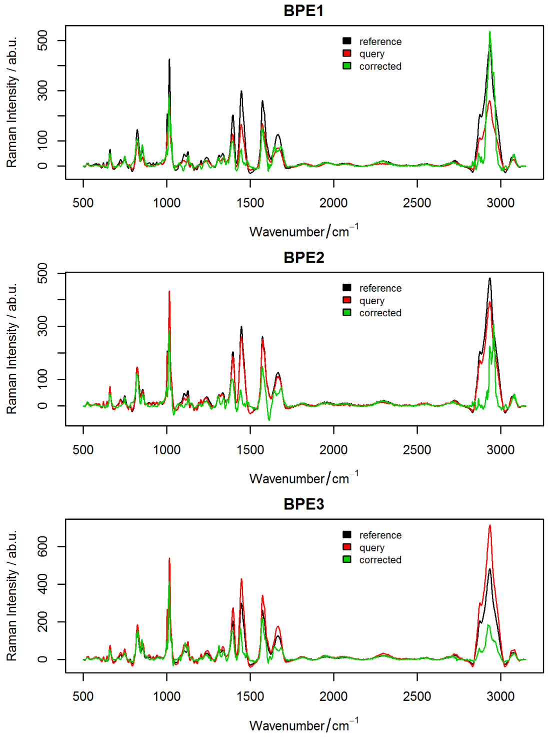


Fig. S7 Average of the reference spectra, average of the query spectra, and average of spectra after PDS for each device.

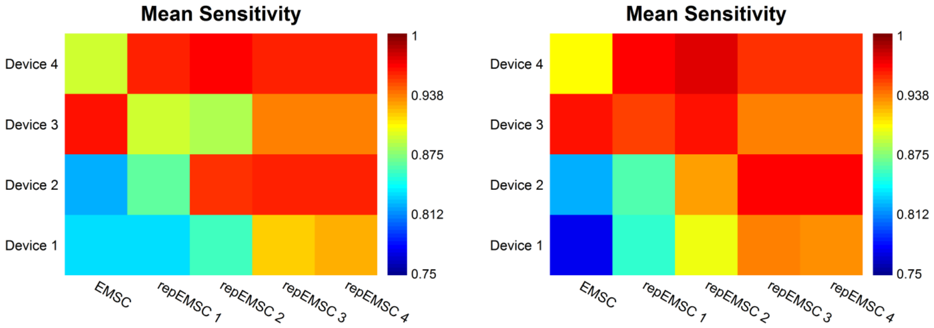
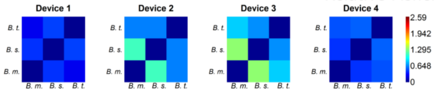
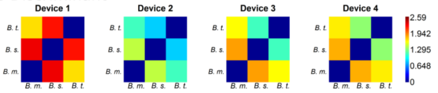


Fig. S8 Comparison of the mean sensitivities for the prediction of each independent device, if different reference spectra were used for the EMSC models. (left) The mean spectrum was used as the reference spectrum, with no baseline correction; (right) the mean spectrum was used as the reference spectrum, with a baseline correction. The latter procedure gave slightly better results as the former case.

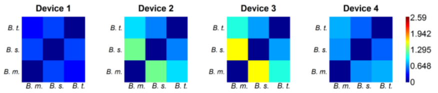
Relative Fisher's Discriminant



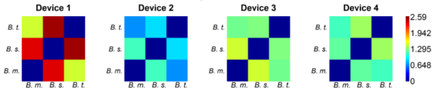
(**a**) Baseline Correction by R^{12}



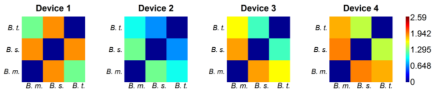
(**e**) Replicate EMSC with Baseline Corrected Reference



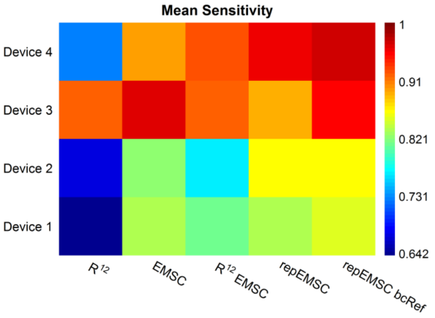
(**b**) Baseline Correction by Basic EMSC



(**c**) Baseline Correction by R^{12} + Replicate Correction

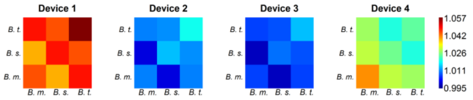


(**d**) Replicate EMSC

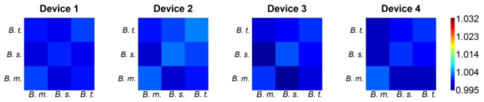


(**f**) Mean Sensitivity of PLSR with LODCV

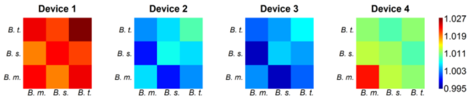
Fig. S9 Values of relative Fisher's discriminant ratio for Raman spectra corrected with different pre-processing procedures. The mean sensitivities of the classification with the same pre-processing approaches were visualized for comparison in subpanel (f). The increasing relative Fisher's discriminant ratios indicate a decrease of inter-replicate deviations (a-e). This was consistent with the improved mean sensitivities of the prediction plotted in false-colors (f). Therefore, the defined markers can be used to evaluate the relative inter-replicate deviations without training a classification model.



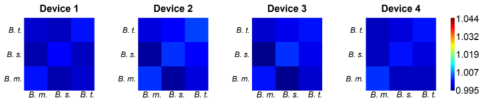
(**a**) Baseline Correction by R^{12}



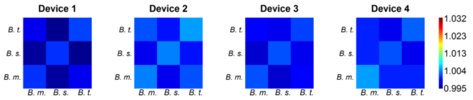
(**d**) Replicate EMSC



(**b**) Baseline Correction by Basic EMSC



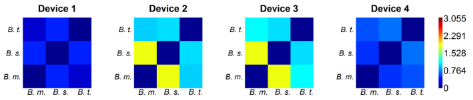
(**e**) Replicate EMSC with Baseline Corrected Reference



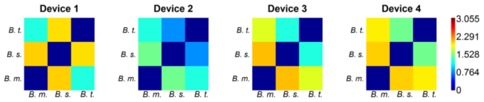
Relative Correlation Coefficient

(**c**) Baseline Correction by R^{12} + Replicate Correction

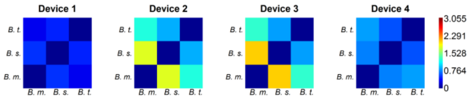
Fig. S10 Results of relative Pearson's correlation coefficient, with Raman spectra corrected by different pre-processing procedures: baseline correction by R^{12} optimization (a) or basic EMSC (b), R^{12} optimization combined with replicate correction (c), replicate EMSC (d), and replicate EMSC using baseline corrected mean spectrum as reference (e).



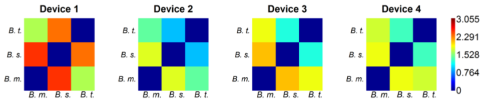
(**a**) Baseline Correction by R^{12}



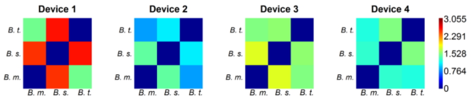
(**d**) Replicate EMSC



(**b**) Baseline Correction by Basic EMSC

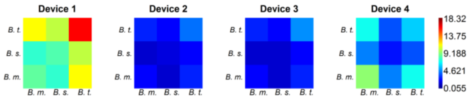


(**e**) Replicate EMSC with Baseline Corrected Reference

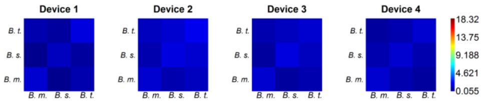


(**c**) Baseline Correction by R^{12} + Replicate Correction

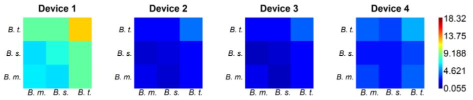
Relative Fisher's Discriminant
from Scores of the PCA



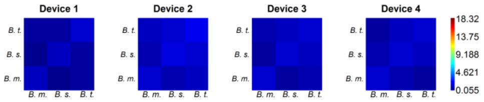
(**a_1**) Baseline Correction by R^{12}



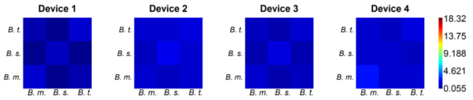
(**d_1**) Replicate EMSC



(**b_1**) Baseline Correction by Basic EMSC



(**e_1**) Replicate EMSC with Baseline Corrected Reference



(**c_1**) Baseline Correction by R^{12} + Replicate Correction

Relative Correlation Coefficient
from Scores of PCA

Fig. S11 Results of the relative Fisher's discriminant (a-e) and relative correlation coefficient (a_1-e_1). The calculation was conducted in different cases of pre-processing: baseline correction by R^{12} optimization (a, a_1) or basic EMSC (b, b_1), R^{12} optimization combined with replicate correction (c, c_1), replicate EMSC (d, d_1), and replicate EMSC using baseline corrected mean spectrum as reference (e, e_1). A PCA was performed on the Raman spectra after pre-processing. The markers were calculated on the scores of the PCA.

Peer-Reviewed Publications

1. T. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler and J. Popp, Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? *Analytical Chemistry*, 2016, **88**(1): 133-151.
2. V. Kumar B.N., S. Guo, T. Bocklitz, P. Rösch and J. Popp, Demonstration of carbon catabolite repression in naphthalene degrading soil bacteria via Raman spectroscopy based stable isotope probing. *Analytical Chemistry*, 2016, **88**(15): 7574-7582 .
3. S. Guo, T. Bocklitz, and J. Popp, Optimization of Raman-spectrum baseline correction in biological application, *Analyst*, 2016, **141**: 2396-2404.
4. S. Guo, R. Heinke, S. Stöckel, P. Rösch, T. Bocklitz and J. Popp, Towards an improvement of model transferability for Raman spectroscopy in biological applications, *Vibrational Spectroscopy*, 2017, **91**: 111-118.
5. S. Guo, T. Bocklitz, U. Neugebauer and J. Popp, Common mistakes in cross-validating classification models, *Analytical Methods*, 2017, **9**: 4410-4417.
6. O. Chernavskaia, S. Guo, T. Meyer, N. Vogler, D. Akimov, S. Heuke, R. Heintzmann, T. Bocklitz and J. Popp, Correction of mosaicking artefacts in multimodal images caused by uneven illumination, *Journal of Chemometrics*, 2017, **31**(6): e2901.
7. S. Guo, S. Pfeifenbring, T. Meyer, G. Ernst, F. von Eggeling, V. Maio, D. Massi, R. Cicchi, F. S. Pavone, J. Popp, and T. Bocklitz, Multimodal image analysis in tissue diagnostics for skin melanoma, *Journal of Chemometrics*, 2018, **32**(1): e2963.
8. S. Guo, R. Heinke, S. Stöckel, P. Rösch, J. Popp, and T. Bocklitz, Model transfer for Raman spectroscopy based bacterial classification, *Journal of Raman Spectroscopy*, 2018, **49**(4), 627-637.
9. S. Guo, O. Chernavskaia, J. Popp, and T. Bocklitz, Spectral reconstruction for shifted-excitation Raman difference spectroscopy (SERDS), *Talanta*, 2018, **186**, 372-380.
10. S. Guo, A. Kohler, B. Zimmermann, R. Heinke, S. Stöckel, P. Rösch, J. Popp, and T. Bocklitz, *Analytical Chemistry*, 2018, **90**, 9787-9795.

Conferences

Talks

1. S. Guo, T. Bocklitz, and J. Popp, *New approach for SERDS spectral reconstruction*, Conferentia Chemometrica 2017, Budapest, Hungary, 3-6, September, 2017.
2. S. Guo, S. Pfeifenbring, T. Meyer, G. Ernst, F. von Eggeling, V. Maio, D. Massi, R. Cicchi, F. S. Pavone, J. Popp, and T. Bocklitz, *Multimodal image analysis for tissue diagnosis of skin melanoma*, Winter Symposium on Chemometrics (WSC 11), St Petersberg, Russia, 26. Feburary-2. March, 2018.

Posters

1. S. Guo, R. Heinke, T. Bocklitz, and J. Popp, *Model transfer problems in biological Raman spectroscopy*, Dokdok 2015, Eisenach, Germany, 11-15, October, 2015.
2. S. Guo, R. Heinke, T. Bocklitz, and J. Popp, *Model transfer for Raman spectroscopy in biological applications*, XVI Chemometrics in Analytical Chemistry, Barcelona, Spain, 6-10, June, 2016.

Workshops

1. *Research data management*, Graduate Academy, Universität Jena, 29-30/04/2015.
2. *Public speaking for scientists*, Graduate Academy, Universität Jena, 06/05/2015.
3. *Speech and vocal training*, Graduate Academy, Universität Jena, 30-31/10/2015.
4. *Scientific presentations*, Graduate Academy, Universität Jena, 09/11/2015.
5. *Good scientific practice*, Graduate Academy, Universität Jena, 17-18/05/2016.
6. *Teaching natural science in higher education*, Graduate Academy, Universität Jena, 21/06/2016.
7. *Teaching natural science in higher education*, OSA Chapter Jena, 09/03/2017.

Acknowledgement

I would like to thank Prof. Dr. Jürgen Popp for giving me the opportunity to carry out my Ph.D study in this highly interdisciplinary working group and for all his intensive support.

I would like to express my great gratitude to PD Dr. Thomas Bocklitz. I am grateful that he is always there for discussions, is open to listen to my ideas, and offers many opportunities for my improvement. His endless patience, continuous support, invaluable suggestions, and kind encouragement during my Ph.D study are forever appreciated.

My special gratitude goes to Prof. Dr. Achim Kohler and Dr. Boris Zimmermann for their warm host and kind support during my short-term exchange in Norwegian University of Life Sciences. I benefit a lot from all the constructive discussions for my research of replicate EMSC-based model transfer. I am expecially grateful to them for their kind suggestions on the related manuscript. I also thank all members in Prof. Kohler's group for their friendliness.

I sincerely thank Dr. Petra Rösch for all her suggestions and advice on my three publications of model transfer. Great thanks go to Ralf Heinke, Dr. Stephan Stökel, Prof. Dr. Ute Neugebauer, and Dr. Olga Chernavskaia for their measured datasets, which have largely supported my PhD research.

The cooperations with Dr. Vinay Kumar Bangalore Narayana, Dr. Tatiana Kirchberger-Tolstik, Dr. Anuradha Ramoji, Jing Huang, Dr. Anna Mühlig, and Björn Lorenz have largely contributed to my improvement and are highly acknowledged.

I thank Oleg Ryabchykov, Dr. Michael Kühnert, Dr. Evelyn Kämmer, and Vlad Iancu, for their warm welcome to me joining the working group, and their generous help, especially in the early days of my PhD study. I especially thank Oleg Ryabchykov for patiently answering many scientific questions. I equally thank Nairveen Ali, Pranita Pradhan, Mehul Chhallani, and Rola Houhou for their kindness, friendship, help, and the nice constructive working environment.

For the correction of my thesis and the German summary I greatly thank PD Dr. Thomas Bocklitz. My gratitude also goes to Prof. Dr. Michael Schmitt for his kind suggestions on my thesis.

For financial support I acknowledge the China Scholarship Council.

I specially thank my friends in Jena, Dongmei, Guangrui, Jingjing, and Xiaoyang, who have strongly supported me during my stay in Germany. My sincere thanks also go to all my dearest friends in China, they have literally been together with me and made me brave all this time. I thank my beloved parents and brother. They are always loving, caring, supporting, and understanding. Words are far less to express my gratitude. I thank them from the deepest inside of my heart.

Erklärungen

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Name der Verfasserin	Datum	Ort	Unterschrift

Erklärung zu den Eigenanteilen der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation (in die kumulative Dissertation aufzunehmen).

Für alle in dieser kumulativen Dissertation verwendeten Manuskripte liegen die notwendigen Genehmigungen der Verlage ("Reprint permissions") für die Zweitpublikation vor.

Die Co-Autoren der in dieser kumulativen Dissertation verwendeten Manuskripte sind sowohl über die Nutzung, als auch über die oben angegebenen Eigenanteile der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation informiert und stimmen dem zu.

Die Anteile der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation sind in der Anlage aufgeführt (Musterbeispiel).

Name der Promovendin	Datum	Ort	Unterschrift

Ich bin mit der Abfassung der Dissertation als publikationsbasierte, d. h. kumulative, einverstanden und bestätige die vorstehenden Angaben. Eine entsprechend begründete Befürwortung mit Angabe des wissenschaftlichen Anteils der Doktorandin an den verwendeten Publikationen werde ich parallel an den Rat der Fakultät der Chemisch-Geowissenschaftlichen Fakultät richten.

Name Erstbetreuer(in)	Datum	Ort	Unterschrift

Name Zweitbetreuer(in)	Datum	Ort	Unterschrift