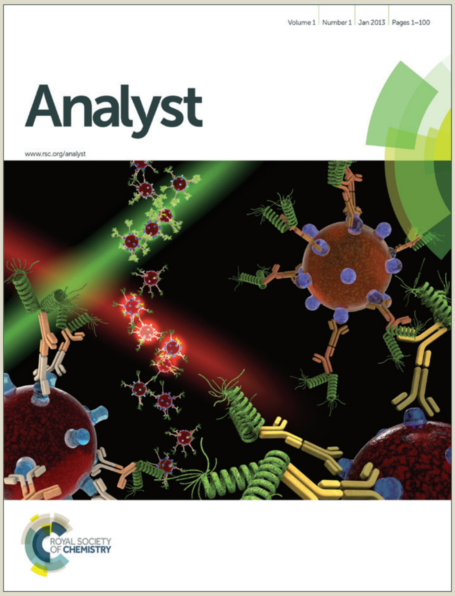
Analyst

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analyst
www.rsc.org/analyst
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Colocalization of Fluorescence and Raman Microscopic Images for the Identification of Subcellular Compartments: A Validation Study[†]

Sascha D. Krauß, Dennis Petersen, Daniel Niedieker, Inka Fricke, Erik Freier, Samir F. El-Mashtoly, Klaus Gerwert, and Axel Mosig[*]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

A major promise of Raman microscopy is the label-free detailed recognition of cellular and subcellular structures. To this end, identifying colocalization patterns between Raman spectral images and fluorescence microscopic images is a key step to annotate subcellular components in Raman spectroscopic images. While existing approaches to resolve subcellular structures are based on fluorescence labeling, we propose a combination of a colocalization scheme with subsequent training of a supervised classifier that allows label-free resolution of cellular compartments. Our colocalization scheme unveils statistically significant overlapping regions by identifying correlation between the fluorescence color channels and clusters from unsupervised machine learning methods like hierarchical cluster analysis. The colocalization scheme is used as a pre-selection to gather appropriate spectra as training data. These spectra are used in the second part as training data to establish a supervised random forest classifier to automatically identify lipid droplets and nucleus. We validate our approach by examining Raman spectral images overlaid with fluorescence labelings of different cellular compartments, indicating that specific components may indeed be identified label-free in the spectral image. A Matlab implementation of our colocalization software is available at http://www.mathworks.de/matlabcentral/fileexchange/46608-frcoloc.

1 Introduction

Identifying overlapping observations between different microscopic images of one and the same sample has been a recurrent topic in microscopic image analysis. While corresponding approaches to identify colocalization patterns between two fluorescence microscopic images are well-established[1,2], there are essentially no established approaches for advanced microscopic setups where samples are measured across different types of microscopes. Yet, cross-microscopy-plattform studies are gaining popularity and relevance. One setting where cross-platform image analysis takes an important role is the combination of Raman microscopy with fluorescence microscopy in order to obtain a label-free protocol to resolve subcellular compartments of cultured cells[3]. A similar setting is found in studies combining other types of vibrational microscopy such as coherent anti-Stokes Raman scattering (CARS)[4] or infrared (IR) microscopy[5] with either fluorescence or brightfield microscopy. In these applications, correlating observations between vibrational spectroscopic images and fluorescence or histopathological staining images is required to obtain training data for supervised classifiers, which allow to resolve compartments of cellular or tissue material without labeling, using only vibrational microscopy.

The main step for the colocalization task is to use fluorescence as a means of "annotation" of spectral images, so that representative reference spectra of different cellular compartments can be collected based on an overlay between a Raman image and a fluorescence microscopic image. These reference spectra can subsequently be used for training a supervised classifier[4] or interpolating contributions of different compartments to an observed location spectrum[3]. Obtaining suitable reference spectra, however, turns out to be a delicate task. A naive approach would be to use spectra from all positions where the fluorescence intensity exceeds a suitable threshold value. However, this would produce a heterogeneous data set for several reasons. This may for instance result from small differences in the z-layer between fluorescence and Raman image, and leads to an imperfect overlay that generally cannot be compensated. Also, differences in confocal volume lead to slight morphological differences between the fluorescence image and the Raman spectral image. To compensate these shortcomings and obtain consistent spectra to train supervised classifiers, one can presegment the spectral image, aiming to identify a segment that has the best possible overlap with the above-threshold positions in the fluorescence image.

In this work, we present a systematic computational ap-

[†] Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

[*] *Department of Biophysics, Ruhr-University Bochum, Universitätsstr. 150, 44780 Bochum, Germany. Tel: +49-234-32-29827; E-mail: axel.mosig@bph.rub.de*

proach to utilize colocalization across different microscopy platforms. This colocalization approach yields supervised classifiers, for which we introduce an appropriate validation measure, which allows us to systematically assess the robustness across a larger set of samples. Our approach utilizes ideas developed in the context of analyzing colocalization between two fluorescence images. Based on presegmentations, our colocalization procedure naturally carries to constellations involving other combinations of microscopes.

Our reference application of resolving the subcellular organization of cells is an important foundation for studying the function of proteins, with applications ranging from identifying disease related location patterns[6–8] to the characterization of drug response[9]. While the gold standard for identifying cell organelles is fluorescence microscopy[10], label-free approaches based on Raman[3] or CARS[4] microscopy promise to overcome the need for fluorescently labeling of the sample under consideration. In this contribution, we present a systematic validation of such *colocalization* studies between vibrational microspectroscopic and fluorescence microscopic images. While one variant of this method has been investigated previously[4], the present contribution provides a more general approach to colocalization involving different colocalization measures including a quantitative comparison of these measures. As a guiding example for our study, we investigate the fully automated identification of nuclei and lipid droplets (LD) in colon and pancreatic cancer cell lines. The knowledge about these two organelles is valuable, because their size, morphology, and amount can be signs of cancer and infections[11–14].

1.1 Segmentation of Raman Microscopic Images

Raman microscopy allows to characterize cell or tissue samples with a pixel resolution of few hundred nano meters, where each pixel location is represented by a Raman emission spectrum. Biologically or chemically relevant information is commonly obtained by high dimensional data analysis of the pixel spectra using techniques such as supervised and unsupervised learning or factorization methods.

Using Raman (and also CARS) microscopy to resolve different parts of subcellular architecture has proven successful in several studies[3,4,15–17], based on a large choice of either clustering approaches or interactive segmentation tools[18]. In order to obtain cellular images from the pixel spectra of a microspectroscopic image, Miljković *et al.*[16] compare methods that segment the pixel spectra of one dataset into base classes, and categorize the commonly employed approaches into *crisp clustering* where each pixel is assigned one similarity class, and *soft clustering* where each pixel spectrum is decomposed into a mixture of several base spectra. Remarkably, the study by Miljković *et al.*[16] as well as most other studies investigate *unsupervised* approaches in the sense that the observed spectra of one dataset are partitioned into base classes. Which of the identified base classes corresponds to which cellular compartment is then left to essentially subsequent visual inspection, e.g. using fluorescence images of the same sample.

The first studies to shift from this unsupervised paradigm to *supervised* approaches are provided by Klein *et al.*[3] and Bocklitz *et al.*[19]. Klein *et al.*[3] systematically overlay a Raman spectral image with fluorescence labelings of the same sample. As each organelle to be identified is labeled by one marker protein, they identify Raman spectral bands that are most informative for one particular organelle by measuring mutual information between spectral bands and fluorescence intensities. These spectral bands are utilized in a supervised learning spirit to infer a nonlinear interpolation function, which can predict a fluorescence intensity from a given pixel spectrum. This results in an intensity image in the spirit of a soft clustering approach. Compared to unsupervised soft clustering, and due to the supervised approach of inferring a prediction function, the base class intensities can be assigned to one cellular organelle. Furthermore, supervised approaches were recently used to automatically identify colon tissue types including adenocarcinoma in Raman spectral datasets[20], following an annotation-based approach as it is commonly employed in IR microscopy based spectral histopathology[5,21]. In the latter studies, random forests (RF) turned out to be convenient tools for supervised classification of both Raman and IR spectra due to their simplicity and efficiency as well as their robustness against overfitting.

While in a previous contribution[4] a colocalization approach was introduced to train supervised classifiers for resolving subcellular architecture, our present contribution provides a systematic comparison between different correlation measures for this approach, along with a cross-validation scheme that provides a more realistic assessment of the classification power than conventional cross-validation. As further contributions of this work, we demonstrate that the colocalization based training of supervised classifiers originally proposed for CARS data in the above mentioned work also performs on Raman spectral images, and assess classifiers for Raman spectra, in particular with respect to different factors such as subcellular organelle, cell type, and confounders.

Supervised classification for resolving subcellular structures has been broadly investigated on the basis of morphological features extracted from fluorescence images[9,22–24]. An advantage of combining label-free Raman microscopy with supervised classification is that once a supervised classifier has been trained, it can be applied to new datasets to identify organelles without any fluorescence labeling or visual inspection of either spectra or segmentations. At the same time, the accuracy of supervised classifiers can be quantified using well-established methods. A complication introduced by Raman microscopy is that both training and, more importantly, validating these classifiers needs to deal with the presence of hundreds or thousands of feature vectors for each component in each cell (namely one vector for each pixel), whereas fluorescence microscopic images yield a single feature vector for each cell and each fluorescence labeling. To deal with the abundance of spectra for training classifiers, our approach follows the procedure typically taken in spectral histopathology to resolve tissue structure in tissue sections[5,21]. In these approaches, one first collects training spectra that are representative for different tissue components. Then, based on these spectra, a supervised classifier is trained. For the validation of Raman spectral classifiers, we use the concept of *leave-one-sample-out* cross-validation, where all spectra from one sample are assigned to either training or validation set. This validation scheme facilitates a systematic assessment of the robustness across a larger set of samples, whereas validation in previous studies was either limited to a single sample[19] or a small number of samples[3], lacking a comprehensive validation measure.

As any supervised classification task, our approach involves recruiting training data, which indeed constitutes the core of our methodological approach. To obtain representative training spectra for different cellular compartments, we overlay the spectral image with its fluorescence counterpart and perform a certain *colocalization analysis*. For this colocalization analysis, we employ ideas that have been extensively and successfully utilized to determine and quantify colocalization between two fluorescence images in previous studies[1,2,25–27]. In our setting, one of the two fluorescence images is replaced by a presegmented version of the spectral image. As it is initially unclear what presegmentation of the spectral image will resolve a particular cellular compartment, we systematically utilize the hierarchy yielded by hierarchical cluster analysis (HCA), as illustrated in Fig. 1. Our approach to identify representative spectra for one cellular compartment in fact reads as identifying a branch in the HCA that exhibits the highest degree of colocalization with the corresponding fluorescence image.

1.2 Colocalization Schemes

In order to quantify which area exhibits the highest degree of colocalization between segments obtained by HCA and a thresholded fluorescence image, we employ colocalization schemes that have been established for measuring colocalization between fluorescence images. Several such approaches have been proposed in the past[1,25,26], as surveyed in Bolte and Cordelieres[2]. Among these measures, the Pearson correlation coefficient (PCC) has gained significant popularity. The PCC is defined as

$$PCC = \frac{\sum_i (R_i - R_{\text{avg}}) \cdot (G_i - G_{\text{avg}})}{\sqrt{\sum_i (R_i - R_{\text{avg}})^2 \cdot \sum_i (G_i - G_{\text{avg}})^2}}, \qquad (1)$$

where R_i denotes the intensity of the first color channel (red) at position i, and R_{avg} the average intensity of the red channel; correspondingly, G_i and G_{avg} represent the pixel and average intensities for the second color channel (green).

This motivates us to introduce the following procedure: For every possible combination of a cluster and a color channel, the degree of colocalization is calculated according to the PCC (see Fig. 1). As every possible cluster from every level of the dendrogram is checked for colocalization, the clusters with the highest PCC found for the two or three color channels might overlap, which means that they are sub- or supernodes of each other. If this is the case, only the one with the highest value is kept in this round and for the remaining color channels a new cluster has to be found.

Note that the first cluster chosen may cover a large area of the image, which may be much larger than the area covered by fluorescence foreground. This may prevent the identifcation of best matching clusters for other organelles. When assessing the suitability of a colocalization measure, it will thus be of crucial importance to determine the number of unidentifiable clusters, which should be as small as possible for an appropriate measure.

2 Methods

2.1 Experimental Materials and Methods

2.1.1 Cell culture. Human pancreatic cancer cells MIA PaCa-2 (CRl-1420) as well as human colon adenocarcinmoa cells HT29 (HTB-38) were obtained from the American Type Culture Collection (kindly provided by Stefan Hahn's laboratory at Ruhr University Bochum). They were treated as described previously[4].

2.1.2 Confocal Raman microscopy. Raman hyperspectral data sets were acquired using a confocal Raman microscope (Alpha300AR, WITec Inc., Ulm, Germany) coupled to a frequency doubled solid state laser operating at 532 nm (Nd:YAG, max. 40 mW, Reno, USA), using a laser power of 10 mW. A 25 μm diameter single-mode optical fiber was used to couple the laser radiation into a Zeiss microscope. The incident laser beam was collimated via an achromatic lens and passed through a holographic band-pass filter before being focused into the sample through a 60x/1.00 NA water immersion objective (Nikon, Japan). The Raman scattered light is collected with the same objective and passed through a holographic edge filter onto a multi-mode optical fiber (50 μm diameter) to a spectrometer equipped with a back-illuminated electron multiplying charge coupled device (emCCD) camera

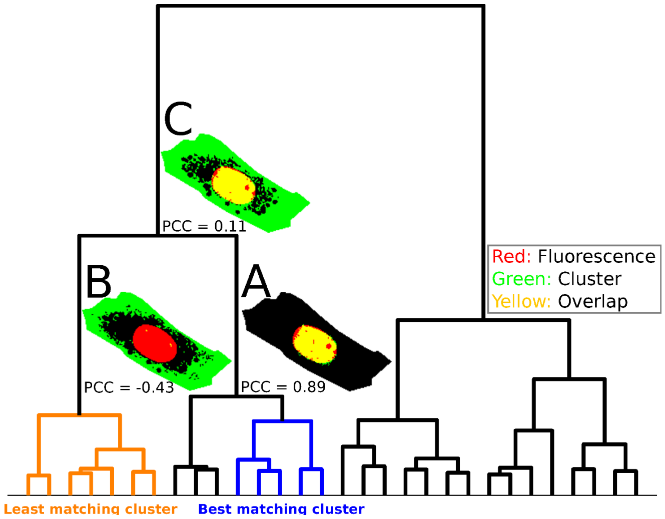


Fig. 1 *Dendrogram from HCA including exemplary overlays of clusters with the fluorescence color channel representing the nucleus.* In the cell overlay plots the overlay of cluster and fluorescence is shown in yellow, the rest of the fluorescence in red, the rest of the cluster in green and the rest of the cell (neither cluster nor fluorescence) in black. *(A)* The best matching cluster (colocalized with a PCC of 0.89) shown in a cell overlay plot and labeled in blue in the dendrogram. *(B)* The least matching cluster (anti-colocalized with a PCC of -0.43) shown in a cell overlay plot and labeled in orange in the dendrogram. *(C)* The last cluster consisting of both the best and least matching clusters, barely colocalized with a PCC of 0.11.

(1600 x 200 px) operating at -60°C. The sample was located on a piezoelectrically driven scanning stage. Raman data sets were obtained by raster-scanning with a pixel size of 0.5 μm for regions of around 60 μm × 60 μm and exposure time of 0.3 s per pixel.

2.1.3 Fluorescence staining and imaging. After permeabilization with 0.2% Triton X-100 for 5 min at room temperature, the cells were washed with PBS and blocked with 1% bovine serum albumin for 30 min. The cells were incubated for 10 min with LD540 (4,4-difluoro-2,3,5,6-bis-tetramethylene-4-bora-3a,4a-diaza-*s*-indacene), washed with PBS-buffer and incubated with 1,5-bis[2-(di-methylamino)ethyl]amino-4,8-dihydroxyanthracene-9,10-dione (DRAQ-5; Cell Signaling Technology, Danvers, USA). The excess fluorescence dyes were removed by PBS-buffer. The fluorescence measurements were performed all the time sequentially on double stained specimen with a confocal laser scanning microscope (Leica TCS SP5 II) using a Leica HCX IRAPO L (25x / 0.95 W) water immersion objective. In order to enable an optimal match with Raman images, stacks of fluorescence images were recorded and the distance between each layer was 0.5 μm.

2.2 Algorithms and Data Analysis

2.2.1 Preprocessing. Cosmic spikes were removed by impulse noise filter[28] and the spectra were interpolated to a reference wavenumber scale. Further data analysis was performed on the normalized data in the region between 700 cm^{-1} and 1800 cm^{-1} and between 2600 cm^{-1} and 3100 cm^{-1}. Spectra from each image data set were hierarchically clustered based on Ward's algorithm using Pearson's correlation distance to obtain a dendrogram.

The fluorescence images were scaled, clipped and manually registered to the spectral images.

2.2.2 Colocalization Scheme. After hierarchical clustering, each branch in the dendrogram is associated with one area in the spectral image comprising a group of similar spectra. For each branch in the dendrogram, a colocalization index with the foreground locations of each corresponding fluorescence image was computed using PCC. The branch exhibiting the highest colocalization index was considered the *best matching cluster*, as formally defined in Supplement 4.1.

Training spectra were extracted from the best matching cluster based on several post-processing steps, aiming on a restriction of the training spectra area to a "condensed" core region. First, 100 intensity thresholds were tested on the fluorescence images ranging from 1% to 100% intensity. The image was binarized by each of these thresholds, and the PCC computed. The threshold achieving the highest PCC was kept as the best colocalizing threshold. In other words, the HCA is also utilized in order to find an optimal fluorescence threshold for each fluorescence channel. With the binarized version of the fluorescence and the best matching clusters, additional enhancements are possible, starting with a connected components filter: The number of nuclei in the image was given and it is tested whether reducing the number of connected components (keeping the biggest ones) to the number of nuclei alters the degree of correlation (without deleting more than half of the pixels). Then isolated pixels are filtered out by grain filtering. Finally, lipid droplets were identified by their specific marker band at wavenumber 1750 cm^{-1}, and masked out whenever they were not covered by corresponding fluorescence foreground.

2.2.3 Implementation. All data processing was implemented in MATLAB Version 8.2 along with the Image Processing and Statistics toolboxes (The MathWorks, Natick, MA).

3 Results and Discussion

3.1 Comparison of Correlation Coefficients

We compared the values of Pearson correlation coefficient (PCC)[25], Mander's overlap coefficient (MOC)[1], intensity correlation quotient (ICQ)[29], and mutual information (MI)[3] on a series of synthetic images involving two color channels (referred to as *red* and *green*, respectively). The image series starts with 0% overlap between the red and the green channel, and overlap between the channels was gradually increased to 100%, see Supplementary Video 1 for an illustration. To illustrate the effect of varying overlap on the different coefficients, binary images were used, while relative intensities result in a similar pattern (Data not shown). The results of these coefficients are plotted against the percentage of red pixels overlapping with green pixels (see Fig. 2 *A*). The ratio of background versus foreground pixels is 1:1, which leads to the desired effect that every coefficient ranges from its minimal to its maximal possible value. While for the PCC and the ICQ, a negative value indicates anti-colocalization, the MOC has no corresponding anti-colocalization indicator as it yields only positive values, which are identical to the percentage of overlap. The MI is also limited to positive numbers. Furthermore, the image series demonstrates a more severe disadvantage of MI, namely that it does not differentiate between the overlap of foreground and background pixels. In other words, the same MI value is obtained for the same degree of colocalization and anti-colocalization.

In a second series of synthetic images, the ratio of background versus foreground pixels was increased to 1000:1 (see Fig. 2 *B*). This high proportion of background pixels, which is realistic as far as small organelles inside cells are concerned, produces very high ICQ values and very low MI values, making them uninformative. The PCC, however, is sensitive to this ratio, whereas MOC does not adapt at all when changing the ratio, as it does not consider the probability of the colocalization.

To confirm these findings on non-synthetic data, we investigated an additional set of 75 Raman microscopic images with fluorescence counterparts. Beside the nucleus, the corresponding fluorescence images label two further organelles, including 29 measurements with a combination of the endoplasmic reticulum (ER) and Golgi apparatus, 13 with ER and mitochondria, 9 with Golgi and peroxisomes, 4 with mitochondria and peroxisomes, and 20 with Golgi and mitochondria. These membrane-rich organelles were used here instead of lipid droplets as they are more challenging to differentiate[4] due to their strong functional and physical connection. Therefore, they are better suited than the more regular morphological (and also spectral) patterns of lipid droplets to demonstrate the differences of the four measures.

As it turns out, the differences between colocalization measures are reflected by the number of samples in which it was not possible to collect training data for at least one of the labeled organelles represented by a fluorescence color channel, because the better matching organelles did not leave enough unmatched area in the HCA for the lesser matching ones. For

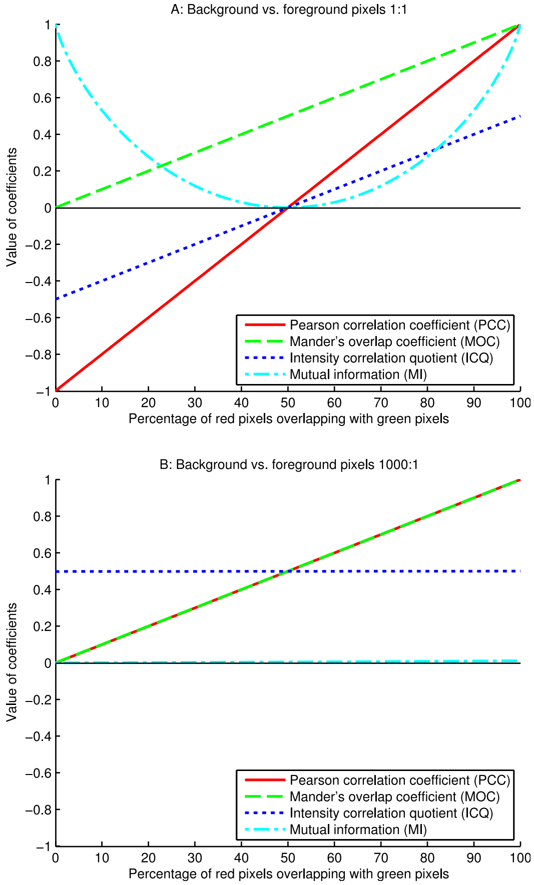


Fig. 2 *Comparison of four different correlation coefficients.* The values of Pearson correlation coefficient (*solid*), mutual information (*dashed*), Mander's overlap coefficient (*dotted*), and intensity correlation quotient (*dash-dotted*) are plotted against the percentage of red pixels overlapping with green pixels. *(A)* The ratio of background versus foreground pixels is 1:1. It can be seen, that only PCC and ICQ indicate anti-colocalization. *(B)* The ratio of background versus foreground pixels was increased to 1000:1 for a further test. This high proportion of background pixels produces very high ICQ values and very low MI values, making them uninformative. PCC adapted to the new ratio, whereas MOC does not change, as it does not consider the probability of the colocalization.

PCC, this was the case in 29 out of the 75 images (38.6%). On using the MI, this number of organelles without training data rose to 43 (57.3%), with the ICQ it was 50 (66.7%) and for the MOC even 61 (81.3%).

This issue is more or less pronounced for different combinations of organelles, where the worst case occurs for ER with Golgi, which are both parts of the endomembrane system. Here, in 55.2% of the images training data cannot be found using the best method (PCC), while it is even 93.1% using MOC. These numbers can be explained by the average size of the best matching clusters found by the different measures: 625 pixels by PCC, 652 by MI, 666 by ICQ and 729 by MOC. On average the biggest cluster (the nucleus) is four times bigger than the next biggest organelle when identified by the PCC, but six times bigger when selected by the MOC. Identifying too large areas as the best matching cluster for the organelle affects identification of best matching clusters of the smaller organelles, as the matching clusters already occupy a (too) large area for the nucleus cluster. This problem becomes particularly obvious for MOC, as its value is determined only by the amount of overlap without taking into account background at all. This behavior favors the identification of larger overlapping areas than the PCC does, where a simultaneous reduction of the two overlapping areas (while keeping the percentage of overlap) increases the value of the coefficient, while it does not change the MOC. While this property of MOC may be desirable under other circumstances, it is inadequate in the context of determining best-matching clusters.

It is important to notice that the effect of unidentified best matching clusters is not represented in the validation of supervised classifiers, as no training data to be (mis-)classified will be contributed to the training data set. This implies that when assessing the quality of a colocalization-based classifier, the number of unidentified clusters for each class is an important quality indicator.

Overall, our observations on both real data and the two synthetic image series clearly support the PCC coefficient as the method of choice to determine colocalization in this work.

3.2 Supervised Classification of Cell Images

Subsequent to identifying best matching clusters for all fluorescence channels, these clusters were used to extract representative training spectra for training a supervised classifier. As shown in Fig. 3 *A-D*, the colocalization provides a best matching cluster for every organelle, in this case the nucleus in blue and the lipid droplets in red. By superimposing these clusters with the corresponding fluorescence color channels an area of overlap appears (shown in yellow), which is the main goal of this procedure: Using this as a mask to recruit the underlying spectra from the Raman image produces a relatively homogeneous data set. The mean spectra of the training data sets for lipid droplets, nucleus and the rest class (consisting of all remaining organelles and the cytoplasm) are shown in Fig. 4. The spectra gained from the colocalization method are used as training data for a random forest classifier[30] using 300 trees.

Note that in Fig. 3, the best matching clusters as well as the agreement between random forest versus fluorescence based segmentations are indicated by their PCC. In order to additionally assess the statistical significance, we computed *p*-values

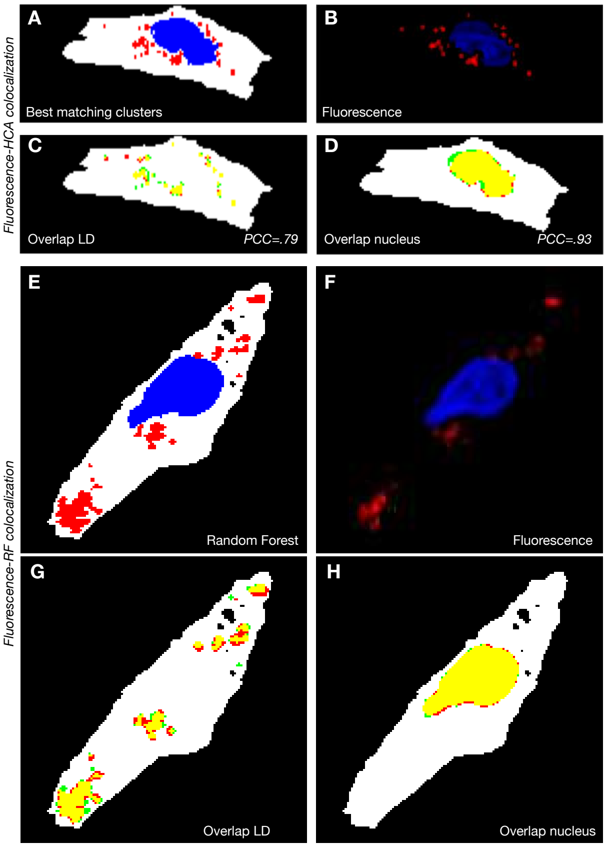


Fig. 3 *Colocalization of Fluorescence with HCA and Random Forest. (A)* The best matching clusters for nucleus (*blue*) and lipid droplets (*red*). *(B)* The corresponding fluorescence image. *(C)* The overlay (*yellow*) of the LD fluorescence color channel (*red*) and its best matching cluster (*green*), colocalized with a PCC of 0.79. *(D)* The overlay of the nucleus fluorescence color channel and its best matching cluster, colocalized with a PCC of 0.93. *(E)* The false color image produced by the RF trained on the spectra derived from *C&D*. *(F)* The corresponding fluorescence image. *(G)* The overlay of the LD fluorescence color channel and the corresponding RF class, colocalized with a PCC of 0.7. *(H)* The overlay of the nucleus fluorescence color channel and the corresponding RF class, colocalized with a PCC of 0.96.

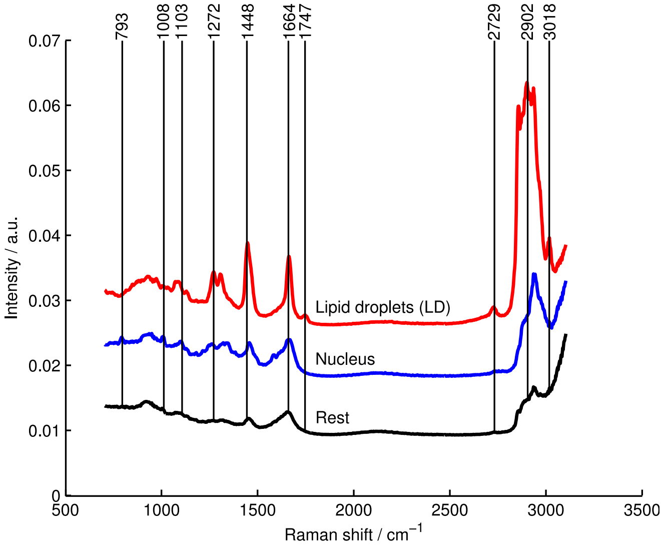


Fig. 4 *Mean spectra of organelles in the training data set.* Lipid droplets, nucleus, and the rest class, consisting of the other organelles and the cytoplasm, are presented. These spectra were automatically collected by the colocalization method to obtain a homogeneous data set for training a random forest.

based on the hypergeometric distribution underlying randomly scrambled pixels as a null hypothesis[26]. For all clusters in our dataset, this *p*-value turns out to be 0, indicating that the correlation is significantly different from randomly scrambled pixels and therefore rejecting the null hypothesis of random overlap.

3.3 Validation of Classification Results

In general, supervised classifiers can be validated in a straightforward manner using different variants of cross-validation such as leave-one-out, *k*-fold, or Monte-Carlo cross-validation. However, in the case of vibrational microspectroscopy, training data are in a sense more structured because each sample contributes not one, but a large number of training spectra for each class. In other words, each class in the training data set is further subdivided into samples (see Fig. 5). In this situation, an important question to be addressed through a suitable validation scheme is whether spectral variability between samples – e.g. due to variability during sample preparation – is a potential confounding factor when classifying for subcellular compartments. Note that this question is generally not addressed by conventional cross validation. To illustrate this, assume an "outlier" sample where all spectra are biased, e.g. through a strong baseline effect affecting all spectra from the sample. As spectra from this same sample will be contained in both the training and the validation set, they can be classified with high accuracy during cross validation. However, in case no spectra from the biased sample are contained in the training data set, classification of the biased spectra will fail during validation.

In order to validate our random forest classifiers appropriately with regard to sample variability, we performed validation using two different approaches (Fig. 5). First, we performed conventional *k*-fold cross validation ($k = 6$) on training data obtained from all six available samples. Next, we performed leave-one-out cross-validation on a *per sample* basis, i.e., the validation set was established from all spectra belonging to one particular sample (see Fig. 5 for an illustration). Both approaches lead to nearly identically high accuracies. In order to simulate high spectral variability between samples, we artificially perturbed all spectra in one of the six samples, and re-evaluated both types of cross-validation. Remarkably, conventional cross-validation was hardly affected by this artifact. While the maximal accuracy of the two versions was identical (100%) and the mean was similar at least (99.5% for *k*-fold vs. 91% for sample-based), the minimal accuracy differed clearly as it was 99.1% for *k*-fold and only 55% for leave-one-sample-out. Compared to conventional cross validation, this indicates that *leave-one-sample-out* provides a more realistic assessment of the quality of a spectral classifier that also assesses spectral variability between samples, as there is no overlap of data from the same measurement between the validation and the training data set.

Nonetheless, the classifier achieved sensitivity and precision values of 97-100% and an accuracy of 99.3% on the original dataset, proving the reliability and consistency of its results and that the colocalization method did produce suitable training data sets. Interestingly, these values for the classifier trained with spontaneous Raman spectral data sets are higher than that of the classifier trained with CARS results[4]. This can be explained in terms of higher spectral resolution of the Raman data sets. In addition, the current Raman spectra provide more spectral information (700-1800 and 2700-3100 cm^{-1}) than that of CARS spectra (2700-3000 cm^{-1}). The data set involving Golgi, ER, peroxisomes and mitochondria achieves a per-sample cross validation accuracy of 91.2%.
Furthermore, the random forest was additionally tested towards its ability to reproduce the results of fluorescence. The degree of correlation between the organelle localization predicted by the random forest and the fluorescence is presented on one of these cells (see Fig. 3 *E-H*), where a high correlation between the results of the two methods can be seen. While in this case a PCC of 0.96 for the position of the nucleus could be observed, the average on 71 cells was 0.86 (standard deviation 0.08). Even when this random forest was tested on a colon cancer cell line (HT29), although being trained on MIA PaCa-2 pancreatic cancer cells, the correlation was at least 0.6 for both organelles on average. This proves the quality of the supervised classifier in reproducing the fluorescence images

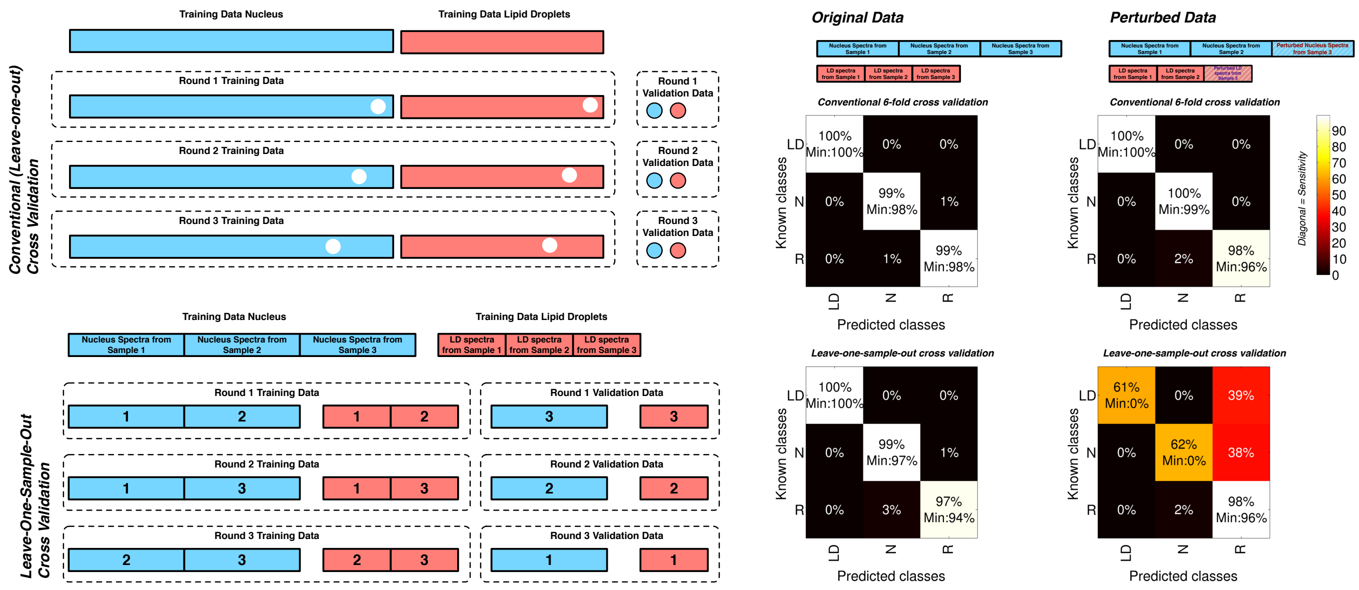


Fig. 5 *Comparison of conventional and sample-based leave-one-out cross validation. Upper Left.* For conventional leave-one-out cross validation each validation round uses one data point from the original training data set for validation, while the remaining data points are used for training. *Lower Left.* In leave-one-*sample*-out cross-validation the validation data set always consists of all spectra from a complete sample, whereas training data are recruited from all remaining samples. *Confusion Matrices.* The confusion matrices are based on six measurements including training spectra of *L*ipid *D*roplets, *N*ucleus and *R*est class. Artificially perturbing spectra in one sample are hardly visible in conventional *k*-fold cross-validation (upper matrices). In leave-one-sample-out validation, however, sensitivities for different organelles are strongly affected (lower right matrix).

without the necessity of using labels or other chemical alterations itself.

Relevance of performing HCA. In order to assess the relevance of using training spectra obtained from the overlap between fluorescence foreground and the best matching cluster, we trained a classifier based on spectra from fluorescence foreground positions for nucleus and lipid droplets, without any utilization of HCA. In this setting, a global threshold in the fluorescence images was determined using the well-established method of Otsu[31]. As it turns out, the accuracy in leave-one-sample-out cross validation drops from 99.3% to 80% in this setting, while the average PCC between the predicted nucleus position and the unthresholded nucleus fluorescence drops from .97 to .62 (refer to Supplementary Figure 5 for an example). This can be explained by mismatches in the fluorescence foreground and the best matching cluster, which seem to be unavoidable und not correctable by registration (see Supplementary Figures 3). Obviously, the utilization of HCA avoids false training spectra in the training data set as also indicated by Supplementary Figure 4 and thus leads to significantly higher accuracy.

3.4 Organelle specificity, cell line specificity, and confounders

Beside the specificity with respect to subcellular organelles, Raman spectra may also distinguish other conditions. To assess this, we trained classifiers to distinguish subcellular organelles of different cell lines. As it turns out, organelles of MIA PaCa-2 cells are spectrally distinguishable from their counterparts in HT29 cells (Supplement 1, classifier *C1*). Furthermore, a classifier may distinguish spectra from the non-cellular surroundings of samples from the two cell lines (classifier *C2*). However, as classifier *C3* indicates, Raman spectra, in particular those observed in areas not covered by any cell, might as well reflect different experimental conditions such as fluctuation of laser power or different laser focus. Yet, the transferability of the organelle classifier between cell lines described above suggests that the spectral differences between different organelles is sufficiently big not to be overshadowed by the spectral differences between cell lines or instrumental conditions. Related phenomena regarding fluorescence signals in non-cellular surroundings have recently been observed for fluorescence markers of subcellular components[32,33].

While subcellular organelles and cell types are biologically relevant factors, spectral classifiers may also *at the same time* distinguish factors that are commonly considered confounders. For example, two different days of experiment can be distinguished in spectra from areas not covered by cells (classifier *C3*). For details on the aforementioned classifiers, we refer to Supplement 1.

4 Conclusion

Our approach extends label-free microscopy for live cell imaging in several directions. It can be seen as the first application of Raman microscopy following a completely supervised paradigm. Furthermore, our approach predicts a crisp segmentation, which makes the result accessible to cross-validation, while soft segmentations are difficult to validate quantitatively. Along the line of quantitative validation, we have shown that sample-based cross validation may uncover problematic effects of spectral variability in the training data and should be preferred as a more realistic assessment of classification power. The results shown in Fig. 5 clearly indicate that leave-one-sample-out cross validation can uncover the usage of unsuitable samples that would have stayed hidden if conventional *k*-fold cross validation had been applied. More generally, leave-one-sample-out validation as a more rigid validity measure may also indicate whether the number of samples in the training data set is sufficient to match the spectral variability between samples. At the same time, the ratio of unidentified best matching clusters for each class should be taken into account when assessing the quality of a classifier. While it would be of interest for future work, currently no objective and quantitative validation scheme for either unsupervised or supervised soft segmentations is available, neither on a *per-spectrum* nor on a *per-sample* basis.

Beside the specificity towards organelles, we could demonstrate that Raman spectra are at the same time specific towards other factors, including factors that are commonly considered as confounders. We also find that Raman subcellular classifiers are transferable (with a loss of accuracy), which has been an issue of investigation recently for fluorescence-based approaches[34]. As the two cell lines under consideration are both epithelial cells, it may need to be answered in the future whether classifiers are also transferable to less similar cell types, for instance stem cells or immune cells. It may also be of future relevance to use our colocalization approach to distinguish cell types, which may be a useful tool for Raman (or CARS) based cell sorting.

Both our present case study of identifying nuclei and lipid droplets, as well as the previous study of identifying other cellular compartments[4] utilizing our novel colocalization scheme, support the claim that colocalization approaches are an important ingredient for obtaining label-free microscopy protocols. Beyond the identification of cellular compartments, colocalization schemes may in general also be useful for resolving tissue structure. In fact, colocalization studies between immunohistologically stained tissue sections and corresponding IR or Raman microscopic images promise a label-free alternative to immunohistochemistry, which is an important tool for tissue diagnostics[35]. Yet, carrying our automated colocalization approach from cells to tissue requires to deal with artifacts of fluorescence microscopy, which are much more pronounced in tissue than they are in cells[36].

Just as the quantitative approaches for colocalization in fluorescence microscopy helped to obtain more reliable conclusions from fluorescence-based studies, our colocalization scheme to align observations between fluorescence and Raman microscopic images promises an objective and highly reproducible approach for label-free microscopy. As correlating observations on one sample across different types of microscopes has gained popularity recently[19,37], colocalization measures provide objective and quantitative means to correlate observations in settings involving other combinations of microscopes.

Finally, utilizing colocalization measures provides further support to utilize hierarchical clustering in a more advanced manner. Conventionally, the dendrogram of hierarchically clustered image spectra is cut “horizontally” to obtain a segmentation into a fixed number of clusters. In Zhong *et al.*[38], however, it has been shown by one of the authors that cutting dendrograms through “non horizontal” cuts yields biologically more meaningful segmentations for IR image spectra. As our newly contributed colocalization scheme generally also identifies such non-horizontal cuts, the present study supports this claim also for Raman spectral image segmentation.

Acknowledgement

The authors thank Ralf Erdmann for providing us with LD540 staining material.

This research was supported by the Protein Research Unit Ruhr within Europe (PURE), Ministry of Innovation, Science and Research (MIWF) of North-Rhine Westphalia, Germany to K. Gerwert and the Center for Vibrational Microscopy (CVM), European Regional Development Fund, European Union and North-Rhine Westphalia, Germany.

References

1 E. Manders, F. Verbeek and J. Aten, *Journal of Microscopy*, 1993, **169**, 375–382.

2 S. Bolte and F. Cordelieres, *Journal of microscopy*, 2006, **224**, 213–232.

3 K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark and J. Schlegel, *Biophysical journal*, 2012, **102**, 360–368.

4 S. F. El-Mashtoly, D. Niedieker, D. Petersen, S. D. Krauß, E. Freier, A. Maghnouj, A. Mosig, S. Hahn, C. Kötting and K. Gerwert, *Biophysical journal*, 2014, **106**, 1910–1920.

5 A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel and K. Gerwert, *Journal of biophotonics*, 2013, **6**, 88–100.

6 D. C. Chan, *Cell*, 2006, **125**, 1241–1252.

7 B. K. Yoder, X. Hou and L. M. Guay-Woodford, *Journal of the American Society of Nephrology*, 2002, **13**, 2508–2516.

8 P. M. McDonough, R. M. Agustin, R. S. Ingermanson, P. A. Loy, B. M. Buehrer, J. B. Nicoll, N. L. Prigozhina, I. Mikic and J. H. Price, *Assay and drug development technologies*, 2009, **7**, 440–460.

9 A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat *et al.*, *Genome biology*, 2006, **7**, R100.

10 M. V. Boland, M. K. Markey, R. F. Murphy *et al.*, *Cytometry*, 1998, **33**, 366–375.

11 M. T. Accioly, P. Pacheco, C. M. Maya-Monteiro, N. Carrossini, B. K. Robbs, S. S. Oliveira, C. Kaufmann, J. A. Morgado-Diaz, P. T. Bozza and J. P. Viola, *Cancer research*, 2008, **68**, 1732–1740.

12 P. T. Bozza and J. P. Viola, *Prostaglandins, Leukotrienes and Essential Fatty Acids*, 2010, **82**, 243–250.

13 Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-F. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell *et al.*, *Science translational medicine*, 2012, **4**, 157ra143–157ra143.

14 S. Yue, J. Li, S.-Y. Lee, H. J. Lee, T. Shao, B. Song, L. Cheng, T. A. Masterson, X. Liu, T. L. Ratliff *et al.*, *Cell Metabolism*, 2014, **19**, 393–406.

15 C. Matthäus, T. Chernenko, J. A. Newmark, C. M. Warner and M. Diem, *Biophysical journal*, 2007, **93**, 668–673.

16 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–2013.

17 S. F. El-Mashtoly, D. Petersen, H. K. Yosef, A. Mosig, A. Reinacher-Schick, C. Kötting and K. Gerwert, *Analyst*, 2014, **139**, 1155–1161.

18 J. Kölling, D. Langenkämper, S. Abouna, M. Khan and T. W. Nattkemper, *Bioinformatics*, 2012, **28**, 1143–1150.

19 T. W. Bocklitz, A. C. Crecelius, C. Matthäus, N. Tarcea, F. von Eggeling, M. Schmitt, U. S. Schubert and J. Popp, *Analytical chemistry*, 2013, **85**, 10829–10834.

20 L. Mavarani, D. Petersen, S. F. El-Mashtoly, A. Mosig, A. Tannapfel, C. Kötting and K. Gerwert, *Analyst*, 2013.

21 B. Bird, S. Remiszewski, A. Akalin, M. Kon, M. Diem *et al.*, *Laboratory Investigation*, 2012, **92**, 1358–1373.

22 M. V. Boland and R. F. Murphy, *Bioinformatics*, 2001, **17**, 1213–1223.

23 N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley and I. G. Goldberg, *Pattern Recognition Letters*, 2008, **29**, 1684–1693.

24 J. Zhou, S. Lamichhane, G. Sterne, B. Ye and H. Peng, *BMC bioinformatics*, 2013, **14**, 291.

25 E. Manders, J. Stap, G. Brakenhoff, R. Van Driel and J. Aten, *Journal of cell science*, 1992, **103**, 857–862.

26 S. V. Costes, D. Daelemans, E. H. Cho, Z. Dobbin, G. Pavlakis and S. Lockett, *Biophysical journal*, 2004, **86**, 3993–4003.

27 J. Adler and I. Parmryd, *Cytometry Part A*, 2010, **77**, 733–742.

28 G. Judith and N. Kumarasabapathy, *Signal and Image Processing: An International Journal, SIPIJ*, 2011, **2**, 82–92.

29 Q. Li, A. Lau, T. J. Morris, L. Guo, C. B. Fordyce and E. F. Stanley, *The Journal of neuroscience*, 2004, **24**, 4070–4081.

30 L. Breiman, *Machine learning*, 2001, **45**, 5–32.

31 N. Otsu, *Automatica*, 1975, **11**, 23–27.

32 L. Shamir, *Journal of microscopy*, 2011, **243**, 284–292.

33 L. P. Coelho, J. D. Kangas, A. W. Naik, E. Osuna-Highley, E. Glory-Afshar, M. Fuhrman, R. Simha, P. B. Berget, J. W. Jarvik and R. F. Murphy, *Bioinformatics*, 2013, **29**, 2343–2349.

34 X. Chen and R. F. Murphy, *Bioinformatics Research and Development*, Springer, 2007, pp. 328–342.

35 P. M. Baker and E. Oliva, *International Journal of Gynecologic Pathology*, 2005, **24**, 39–55.

36 J. Pawley, *Handbook of biological confocal microscopy*, Springer, 2010.

37 R. Masyuko, E. J. Lanni, J. V. Sweedler and P. W. Bohn, *Analyst*, 2013, **138**, 1924–1939.

38 Q. Zhong, C. Yang, F. Großerüschkamp, A. Kallenbach-Thieltges, P. Serocka, K. Gerwert and A. Mosig, *BMC bioinformatics*, 2013, **14**, 333.