Ph. D. Dissertation

Study on single cell Raman analysis to enhance differentiability of cell types in non-homogeneous environments

(不均一環境下における細胞識別性向上に関する1細胞ラマン解析研究)

by

Abdul Halim Bhuiyan (Student ID:28195505)

Under the supervision of

Professor Tamiki Komatsuzaki

Graduate School of Chemical Sciences and Engineering

Hokkaido University, Japan

Friday 18th August, 2023

Acknowledgment

Firstly, I am very much thankful to the Almighty Lord for providing me the strength to attain this achievement of my life. I have accomplished this thesis with the support of many nice people who gave me the encouragement, assistance, and valuable suggestions. I am very much gratitude to all of them. I would like to gratefully thank my supervisor Professor Tamiki Komatsuzaki, for his tremendous guidance, encouragement, expertise, patience, motivation, and cooperation during my entire period of study at Hokkaido University. His incredible ideas provided me with a great opportunity to be learned about the diversify research field that I didn't know ever before. I am very much thankful to Professor Tamiki Komatsuzaki for his contributions of excellent research ideas and funding me to make my PhD research productive and amusing. I would like to thank all the lab members of Professor Komatsuzaki group who were very friendly and supportive. I have learned several things from this group that will be very much helpful for my future professional life. Specially, I would like to express my sincere gratitude to assistant professor Dr. Jean-Emmanuel Clement for supporting me during my PhD research with valuable scientific ideas, many important discussions, encouragement. I have learned from him the concepts of Raman microscopic images and Raman data preprocessing which is the main part of this thesis. I would like to thankful to you Jean-Emmanuel, for your continuous support, time and patient that have been great contributors to accomplish the thesis. I would like to thank Associate Professor Koji Tabata, a very polite and generous person who helped me many times to fix my computer problem, Python coding problem. He is really an expert in Algorithms and coding. I would like to thank ex-Specially Appointed Assistant Professor J. Nicholas Taylor, who is an expert in Raman data analysis. I have learned from him about Raman microscopic images and how to analysis of Raman data by MATLAB. He also supports me by giving constructive suggestions to improve my research. I would like to thank all collaborators of the JST/CREST project, especially, Professors Katsumasa

Fujita, Yoshinori Harada, and Atsuyoshi Nakamura for their valuable comments and discussions. I sincerely thank my sub-advisers Professor Tetsuya TAKETSUGU, Professor Hajime ITO, and Professor Keisuke Takahashi for their valuable comments to flourish the thesis. I would like to express my deep gratitude to Khalifa Mohammad Helal, who encouraged and helped me to enter the PhD program and support me for the initial settlement in my daily life in Sapporo. I would like to thank Udoy Sankar Basak, who receive me at the airport, and for logistic support in Sapporo. I would like to thank all former and current members of our lab without mentioning individual names in person. I spent very wonderful time working with this interesting group and supporting environment. I would like to thank Maiko Ishida ex-secretary and current secretary, Matsue Koida in our lab for their support and translation not only for academic purposes but also for daily life. I gratefully acknowledge all the funding sources especially Advanced Graduate School of Chemistry and Materials Science (AGS) and DX Doctoral Fellowship of Hokkaido University that helped to accomplish my PhD research. I would like to acknowledge Japan Science and Technology Agency (JST)/ Core Research for Evolutional Science and Technology (CREST) for the financial support during PhD research in Japan. I would like to express gratitude to the Bangladesh University of Engineering and Technology, Bangladesh, for allowing me a study leave for the entire period of my research at Hokkaido University, Japan. I would like to express my deep gratitude towards my family members, my source of strength, for their worship and support. I am also very much thankful to my parents for their encouragement, unconditional love, and worship of me to achieve this stage. I would like to express my gratitude to my lovely wife Fatema Khatun for her inspiration and motivational speech during my hard times. It was difficult to accomplish this work without her great sacrifice. I would like to express my thanks for our amazing and invaluable gift, my pretty daughter Abida Tabassum Tuba and son Omar Faruk, who are always the reason for my pleasure in any circumstances of my life.

Abdul Halim Bhuiyan

Hokkaido University, Date: Friday 18th August, 2023

Abstract

Raman imaging is a powerful technique used in biological sample measurement. It gives both spatial and spectral representation of the sample, that can be integrated with machine learning systems to develop new medical diagnosis tool. The Raman measurements were performed with an high-speed Raman microscope, the slit-scanning Raman microscope. It extracts the underlying spatial and spectral information of a sample typically two orders of magnitude faster than raster scanning. In this study, thyroid cell lines, FTC-133(cancerous) and Nthy-ori 3-1(normal) were used as a model to investigate the pertinence of Raman spectroscopy in the diagnosis of thyroid cancer. Line illumination Raman microscope extracts the underlying spatial and spectral information of a sample, typically, a few hundred times faster than raster scanning. This makes it possible to measure a wide range of biological samples such as cells and tissues – that only allow modest intensity illumination to prevent potential damage – within feasible time frame. However, a non-uniform intensity distribution of the laser line illumination may induce some artifacts in the data and lower the accuracy of machine learning models trained to predict sample class membership. Here, using cancerous, and normal human thyroid follicular epithelial cell lines, FTC-133 and Nthy-ori 3-1 lines, whose Raman spectral difference is not so large, I showed the standard preprocessing of spectral analyses widely used for raster scanning microscope introduced some artifacts. To address this issue, I proposed a detrending scheme based on random forest regression, a nonparametric model-free machine learning algorithm, combined with position-dependent wavenumber calibration scheme along illumination line. It was shown that the detrending scheme minimizes the artificial biases arising from non-uniform laser source and significantly enhances the differentiability of the sample states, i.e., cancerous or normal epithelial cells, compared to the standard preprocessing scheme.

Contents

List of Figures

List of Tables

1

Introduction

Raman spectroscopy is a non-invasive technique that has significantly influenced the field of molecular analysis and characterization. Named after Sir C.V. Raman[5]., the Indian physicist who discovered the Raman effect in 1928, this technique works on the principle of inelastic scattering or shift of frequency of monochromatic light, typically a laser, leading to the generation of a unique and detailed spectral pattern of the sample under study, characteristic of its vibrational or rotational state. When compared to other spectroscopy techniques, such as IR, UV-Vis, or fluorescence spectroscopy, Raman spectroscopy offers distinctive advantages. It can analyze complex samples in solvent environments without the need for labeling or injecting any probes. This capability simplifies the sample preparation process and allows for a more straightforward analysis of the material in its natural state. Over the last two decades, these qualities have led to Raman spectroscopy's wide

exploration in the fields of biology and disease diagnosis. The technique's ability to offer detailed molecular insights without significant alterations to the sample holds significant promise for the development of next generation of disease diagnosis tools, including its potential use in "in vivo" scenario, bringing us a step closer to real-time, non-invasive diagnosis. Advances in instrumentation have allowed for the development of Raman microscopes that can scan the spatial dimension of samples. This means that Raman spectroscopy can now be used to create images of samples, as well as spectra. The data collected from these novel microscopes is typically referred to as a hyperspectral Raman image. However, like any technique, Raman spectroscopy also has its own challenges. A key obstacle lies in data standardization. With the absence of universally recognized methods for data acquisition, pre-processing, and analysis, comparing results across different studies becomes a difficult task. Variations in instrument design, sample nature, and data processing approaches can further complicate matters.

1.1 Raman measurement of my research

Raman measurement was used to characterize the vibrational modes of cancer and non-cancer human thyroid cell lines, FTC-133 (cancer) and Nthy-ori 3-1 (non-cancer), respectively. Representative experimental images at specific wavenumbers are shown in Fig. 1.1. Various notable Raman peaks of FTC-133 and Nthy-ori 3-1 cell lines were observed at the different positions of wavenumber shown in the following Table 1.1 [2–4, 6, 7].

A home build line illumination Raman microscopy were used to perform the Raman measurement. Line illumination Raman microscopy uses a laser illumination that is shaped as a straight line and scans the sample from left to right, collecting Raman spectra simultaneously at each spatial position along the line axis (Fig. 1.2). The line illumination axis in the following is denoted as ζ while the scanning axis, which is perpendicular to ζ, is denoted as ξ. The line intensity variation is the primary cause of a non-homogeneous illumination. The line intensity profile often deviates from the theoretical Gaussian profile due to some laser alignment inaccuracies or lens qualities degradation. Subsequently Ra-

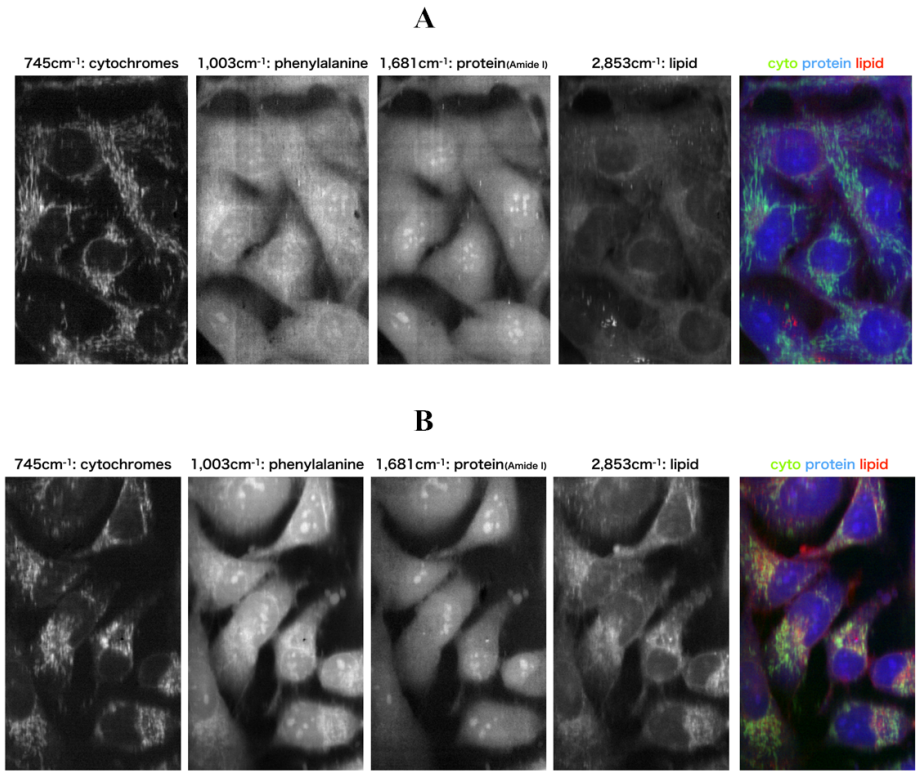


Fig. 1.1. Experimental images of (A) FTC-133(FTC) (B) Nthy-ori 3-1(Nthy).

man spatial distribution at specific Raman shift follow such deviations. An illustration of this deviation is demonstrated in Fig. 1.3, which shows a non-linear intensity profile at 325 cm^{-1}, a prominent peak of calcium fluoride substrate (CaF_2).

Raman peaks	
Wavenumber peak($\pm 3 cm^{-1}$)	Assignment
669	ν_7(δ: porphyrin deformation), observed in the spectra of single human Red Blood Cell (RBC)
720	DNA
750	Cytochromes
811	O-P-O stretching RNA
853	Ring breathing mode of tyrosine & C-C stretch of proline ring Glycogen
956	Crotenoids (absent in normal tissues)
980	C-C stretching β-sheet (proteins)=CH bending (lipids)
1004	Phenylalanine
1076	C-C (lipid in normal tissues)
1048	Glycogen
1127	Cytochromes
1210	Phenylalanine and Tryptophan (Amide III)
1264	Triglycerides (fatty acids)
1307	Cytochromes
1337	Amide III & CH_2 wagging vibrations from glycine backbone & proline side chain A, G (ring breathing modes in the DNA bases) C-H deformation (protein)
1339	Tryptophan
1406	ν_s COO^-(IgG)
1443	CH_2 deformation (lipids and proteins) Triglycerides (fatty acids)
1447	CH_2 bending of proteins and lipids
1490	DNA
1544	Amide II
1584	Cytochromes
1655	Amide I (of collagen)
2850	ν_s CH_2,lipids, fatty acids CH_2 symmetric
2885	ν_s CH_3, lipids, fatty acids
2890	CH_2 asymmetric stretch of lipids and proteins
2913	CH stretch of lipids and proteins
2935	chain end CH_3 symmetric band
3015	ν =CH of lipids

Table 1.1: Assignment of the important peaks in the Raman spectra of FTC-133 and Nthy-ori 3-1 cell lines[2–4].

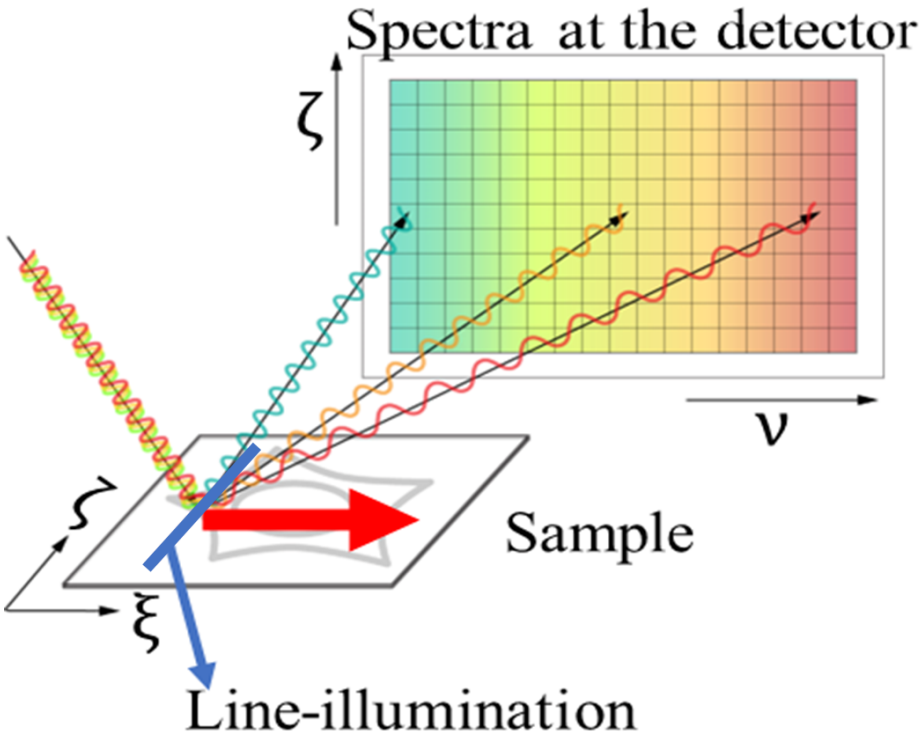


Fig. 1.2. The concept of Line illumination Raman microscope.

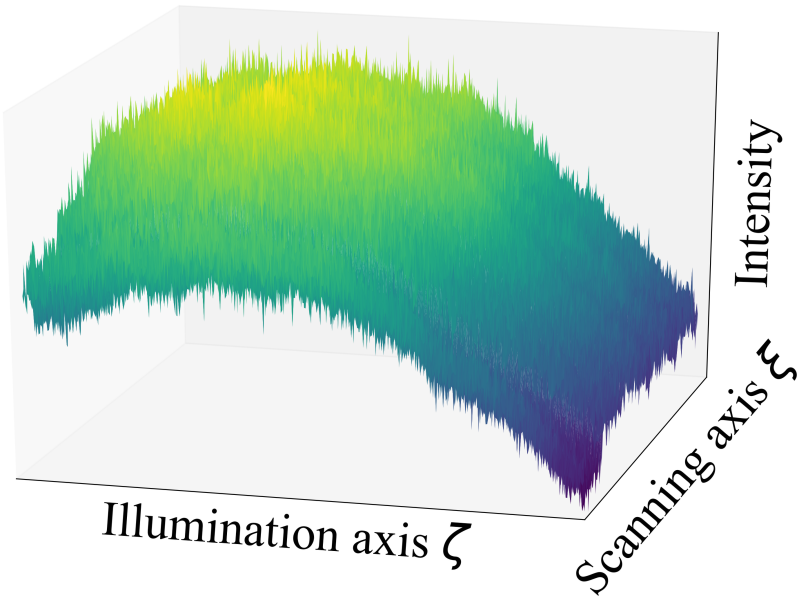


Fig. 1.3. A surface plot of the Raman intensities at 325 cm^{-1} known as a prominent peak common to calcium fluoride (CaF_2) for a representative Raman image FTC-133.

1.2 Literature review

Thyroid cancer starts in the thyroid gland, which is located in the front of the throat, below the larynx. The thyroid gland contains two types of cells: follicular cells and C cells. Follicular thyroid cancer is a challenging cancer to diagnose because most patients do not experience any symptoms, even if they have cancer. Raman data of human thyroid cell lines is used for analysis in this work. Follicular thyroid carcinoma is an invasive and challenging diagnosis to identify malignant form of thyroid cancer based on morphology description[2, 3, 8–12]. However, Raman hyperspectral microscopy is a powerful tool based on an inelastic scattering of light is used in various fields such as chemistry, biomedical and material science etc. which provides the chemical structure and spatial distribution of potential biomarkers in samples of molecules [4, 13].To detect specific molecular anomalies occurs in cancer processes at the cellular level, Raman spectroscopy is very effective tool as it is sensitive to biochemical changes[14]. Raman spectroscopy is a label free, non-destructive technique and sensitive to the molecular changes. It gives a fingerprint of the material. It is a non-invasive diagnostic technique that allow to study living cells.

Deriving meaningful information from raw Raman spectra presents a considerable challenge due to the contamination from noise and extraneous background signals. A dominant factor in this complexity is autofluorescence, which is orders of magnitude greater than Raman scattering, leading to spectra that are predominantly overshadowed by fluorescence. Fluorescence intensity is several orders of magnitude more intense than the weak Raman scattering, Charge-coupled device (CCD) detector is also another source of noise is reported by [1, 15]. So, to remove noise is a very important task before analyzing the data. Background removal during the denoising process were investigated with thresholding strategy using wavelet transform. They found satisfactory results in both simulated and real signals without damaging their shape and area[15]. They applied their approach both simulated and real Raman spectra.

A review paper on label-free brain tumor imaging using Raman-based methods was reported by Hollon et al. [16]. They reviewed the articles on the application of three Raman-

based imaging methods to neurosurgical oncology as well as the machine learning approach. They focused on the improvement of brain tumor patients by detecting tumor infiltration, and guiding tumor biopsy by using label-free Raman-based imaging methods. Martin et al.[17] worked for the development of an advanced hyperspectral imaging system combining several recent advances in photonics technologies, including: a LCTF, an ICCD camera, and a coherent fiber bundle, to produce a viable system for cancer detection. Their system can record fluorescence and white light images at multiple wavelengths rapidly.

Raman spectroscopy is applied to analyze the apoptosis of single human gastric cancer cells inducing 5-FU drug during incubation[6]. They showed that Raman spectroscopy is a sensitive technique for detecting the apoptosis of human gastric carcinoma cells. They performed PCA analysis and conclude that the discrimination achieved was mainly due to scores in PC1. Raman measurement was performed from the cytosol of a living HeLa cell by Palonpon et al.[18]. For live-cell imaging, they developed the Raman spectra that can attain high spatial and temporal resolution. They observed that tiny tags in the cellular silent region of the Raman scattering provide useful information of the chemical specificity to tag target molecules with minimal perturbation.

Classification models are very familiar examples of machine learning algorithms. However, Deep Learning (DL) is a subset of machine learning that contains some complex architecture and achieved promising results compared to conventional machine learning algorithms. DL has been employed with excellent performance to a wide range of computer vision problems for about a decade[19–22]. Particularly, DL have widely used in speech recognition, image classification, object detection, natural language processing, etc. For classification purpose, DL is used for cat and dog images, handwriting digit, fashion dress image, vehicle images and shows tremendous success[23–26].

DL algorithms are mainly developed for image data analysis and several deep neural networks exist for that. But for spectral data analysis, we need to develop DL architecture by tuning hyperparameters and need to think about the amount of data. Because DL is a data-hungry model. More data means the model learns more which is a bit challenging for spectral data. Nevertheless, the researcher has been carried out their research by applying a

data augmentation approach. For data augmentation, some researchers add random noise, shift along the wavenumber axis, and taking a linear combination of the same phenotypes [27, 28].

Deep learning is a very powerful technique compared to machine learning but the challenge is to interpret the results as it is a black box method we don't know exactly what is going to happen inside the architecture. Feature selection is the key tool for machine learning problems as it reduces the computational complexity of the models, and helps for understanding data. In machine learning, we have several feature selection methods that tell us the importance of the features according to the scores. But it is not so much clear how DL can be employed in the feature selection problem. Although during training the model, DL is unable to perform feature selection. That is why we should rethink identifying the important features after the trained model.

Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning is performed by Ho et al.[29]. In this research, they achieved 82% accuracy for low signal-to-noise spectra by deep learning. They reported that high signal-to-noise ratios (SNRs) are needed to reach high identification accuracies. They compared their CNN results with logistic regression (LR) and support vector machine (SVM) and observed significantly better performance of CNN. They also verified their CNN performance by statistical test by two-sample test of sample means. For the LR and SVM, they performed PCA and kept 20 principal components and concluded that using only the first 20 principal components not only decreases computation costs, but also increases accuracy by reducing the amount of noise in the data.

Germond et al.[13] observed that Raman spectral peak intensities significantly correlated with the gene expression of some genes contributing to antibiotic resistance genes. They applied PCA followed by discriminant analysis (DA-PC) for the spectral data classification. They selected eight principal components for their model based on the Fisher score. They performed a statistically significant test and calculated F-value to keep the PCs in the subsequent discriminant analysis. Their proposed model exhibited 100% well-classified observations on the training dataset as well as 100% successfully discriminated

of the test data.

CNN is performed with raw and preprocessed Raman spectra from extracellular vesicles (EVs) to find tumor-derived EVs by Lee et al.[30]. The proposed architecture of the CNN model consists of three sets of convolution layers followed by a max-pooling layer and four hidden layers. To compare the results of CNN, they performed PCA-linear discriminant analysis (LDA) and PCA- quadratic discriminant analysis (QDA) on preprocessed data and raw data. For PCA discriminant analysis, they found good performance on accuracy only preprocessed data. Surprisingly, they found excellent performance on accuracy for CNN on raw data. What are the regions behind this miracle they did not explain in detail? Classification accuracy was better in the fingerprint regions compared to high frequency and full-spectrum regions in their analysis. Residual neural network (ResNet) has been applied to decode Raman spectra-encoded suspension arrays (SAs) by Chen et al.[31]. To visualize their classification performance, a t-distributed stochastic neighbor embedding (t-SNE) was used. Their proposed model gives 100% classification accuracy. They compared their ResNet results with other machine learning classifiers kNN, SVM, FC, and CNN, and found excellent performance over the other models. They used the normalized Raman spectra data without background subtraction in their analysis.

CNN is used to identify components in mixtures by Fan et.al [32]. They confirmed their CNN results with other machine learning methods LR,kNN,RF, and back propagation artificial neural network (BP-ANN). They performed a simulation experiment to identify components in both simulated and real Raman spectral datasets of mixtures and found that their CNN outperformed compared to others. To check the stability of the CNN, they performed one hundred training for the CNN models of acetonitrile and methanol on a simulated test set and found that their distributions of accuracy follow the approximate Gaussian distributions with small variances.

Raman spectroscopy along with ML was used for endometriosis[33]. Their results show that kNN-weighted method was the best classification model with sensitivity and specificity values of 80.5% and 89.7%, respectively. They performed feature selection based on the highest mean accuracy value and selected spectral interval (790–1729 cm^{-1})

then they applied PCA to extract the relevant features for the selected spectral region. They checked the statistical test (Student's t-test) on the data and concluded that there were no statistically significant differences between the patient and control groups in terms of 4 measurements such as age, BMI, uterine myomas, and adenomyosis.

Intraoperative brain cancer detection with Raman spectroscopy in humans was performed by Jermyn et al.[34]. In this research, they applied boosted trees machine learning method to analyze the spectra and checked the performance measure accuracy, sensitivity, specificity, and AUC. They employed a leave-one-out cross-validation approach and analyzed statistically their classification accuracy by performing two-sided normal-based test.

Breast cancer histology images classification using Convolutional Neural Networks was done by Araujo T. et al.[35]. They employed CNN for feature extraction and classification purpose. They performed binary and multiclass classification. Even after feature extraction by CNN the applied SVM for classification.

Raman spectroscopy of breast cancer cell data is analyzed for classification using one-dimensional convolutional neural network by Ma et al. [36]. They performed data augmentation to increase the number of spectra. For the comparison of the performance of their 1D-CNN model, FDA and SVM with ten-fold cross-validation was used to classify two types of breast samples. PCA with 20 components was applied to reduce the dimensionality and complexity of the data set. Finally, they evaluated sensitivity, specificity, and overall accuracy. They found that the performance of CNN model largely depends on the learning rate and batch size during the training process. They observed that 1D-CNN algorithm has higher accuracy than many other common algorithms and is capable to extract features from the input spectra. Also, the combination of Raman analysis and 1D-CNN model may be possible candidate to monitor the therapy of patients which was reported by some other researchers.

Different classification algorithms were used to solve the problem of pigments visualization, classification and identification via Raman spectral[37]. SNR (Signal to noise ratio) is introduced to evaluate the stability of algorithms. They found that for the low

SNR value, the accuracy of the algorithm decreases and for the high SNR value accuracy increased. They performed RF for feature importance and effects of different hyperparameter on the accuracy. Moreover, they checked performance of the classification algorithm after applying the denoising algorithm considering low SNR. Denoising algorithm is very important based on the classification algorithms what they observed.

Raman spectroscopy was used for classification of COVID-19 patients in this paper[38]. They considered 3 cases: COVID-19 patients, suspected cases, and healthy patients and try to distinguish these 3 groups based on classification as well as statistical tests. They performed the pairwise difference between 3 groups. They observed significance difference between COVID-19 patients vs healthy patients in compare to other pairs.

In this thesis, a novel detrending scheme for preprocessing Raman data was developed. The detrending scheme objectives to enhance the quality of Raman spectra by removing some artifacts caused by experimental factors and to enhance the differentiability between two phenotypes.

1.3 Layout of the thesis

The dissertation provides a comprehensive overview of the research conducted in the field of Raman spectroscopy along with machine learning and focuses on the improvement and evaluation of my novel detrending scheme for preprocessing Raman data.

In Chapter 1, the purpose of the dissertation and background of the study have been described.

In Chapter 2, a brief description of some preliminary concepts like Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Random Forest (RF) etc. have been discussed with illustrations.

In Chapter 3, three preprocessing scheme for the Raman data of FTC-133 and Nthy-ori 3-1 cell lines have been emphasized.

In Chapter 4, a comparative study of these three schemes to differentiate two phenotypes, FTC-133 and Nthy-ori 3-1 are explained in detail.

In Chapter 5, all the important results are summarized, and future plans have been described.

2

Preliminary discussions

For the quantitative analysis of the data, several algorithms are needed. In this chapter, I will explain briefly some essential topics/methods that are used in this research.

2.1 Performance Metrics for Machine learning

2.1.1 Confusion matrix

A confusion matrix demonstrates the performance of the classification model. This is one of the techniques to summarize classifier performance. The confusion matrix can be explained as follows:

	Predicted Positive(P)	**Predicted Negative(N)**	
Actual Positive(P)	True Positive(TP)	False Negative(FN)	**Sensitivity** TP/(TP+FN)
Actual Negative(N)	False Positive(FP)	True Negative(TN)	**Specificity** TN/(TN+FP)
	Precision(PPV) TP/(TP+FP)	**Negative Predictive Value(NPV)** TN/(TN+FN)	**Accuracy** (TP+TN)/(TP+TN+FP+ FN)

Fig. 2.1. Confusion matrix

TP: the number of true positives (actual cancer predicted as cancer), FP: the number of false positives (actual non-cancer predicted as cancer), TN: the number of true negatives (actual non-cancer predicted as non-cancer), and FN: the number of false negatives (actual cancer predicted as non-cancer).

F1 score is a weighted average of precision and recall (sensitivity).

$$Recall = \frac{TP}{TP+FN}$$

,

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Accuracy is mainly focused on the correct prediction (True Positives and True negatives)while F1-score is focused on the wrong prediction (False Negatives and False Positives). False Negative is very important for cancer identification. The model tells noncancer but actually, it was cancer, which is very dangerous in medical science.

2.1.2 Receiver Operating Characteristics curve

The receiver operating characteristic (ROC) curve is one of the important curves is used to evaluate the performance of the model which involves the true positive rate (TPR), and the false positive case rate (FPR). TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP+FN}$$

,

$$FPR = \frac{FP}{FP+TN} = 1 - Specificity$$

The area under the ROC curve, known as the area under the curve (AUC) is the quantifier of the performance. The larger the value of AUC means that the performance of the model is better and AUC=1 for perfect classification.

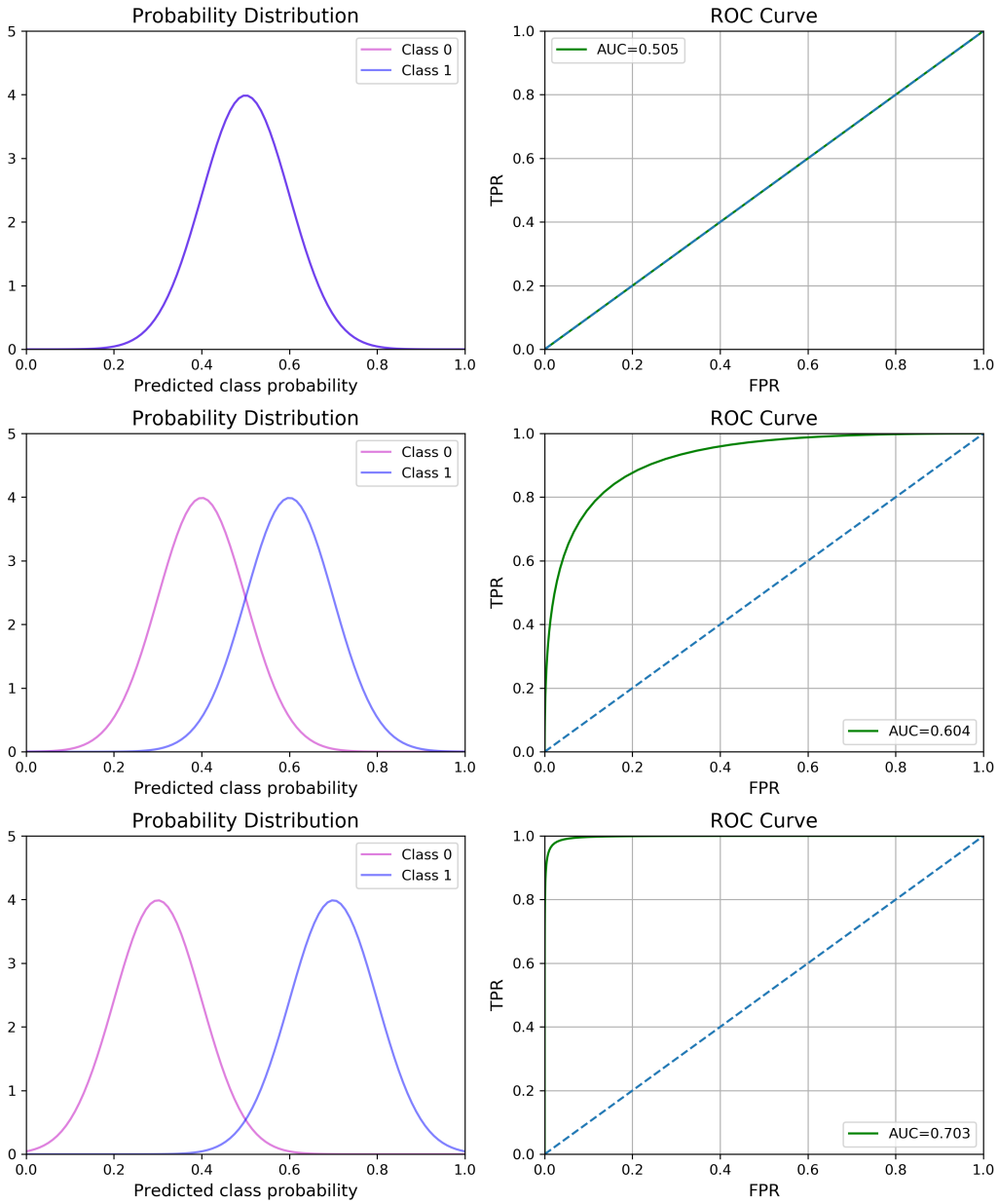


Fig. 2.2. Class distribution and ROC curve

2.2 Principal Component Analysis

Principal Component Analysis (PCA)[39–42] is one of the most important dimensionality reduction and feature extraction methods that obtain important variables from a huge set of variables available in a data set. It extracts a set of features in low dimensional space from high dimensional data by taking a projection of irrelevant dimensions and capturing the data information as much as possible. To perform PCA, we need to know about the covariance matrix and eigenvalue-eigenvector concepts. The first principal component (which corresponds to the largest eigenvalue) is a linear combination of original predictor variables that takes the maximum variance in the data set. I am going to explain PCA with a simple example below. Consider the matrix

$$A = \begin{bmatrix} 126 & 78 \\ 128 & 80 \\ 128 & 82 \\ 130 & 82 \\ 130 & 84 \\ 132 & 86 \end{bmatrix}$$

After transforming the original data to Z-scaled($\frac{x-\mu}{\sigma}$), we get

$$A_{scaled} = \begin{bmatrix} -1.566699 & -1.549193 \\ -0.522233 & -0.774597 \\ -0.522233 & 0 \\ 0.522233 & 0 \\ 0.522233 & 0.774597 \\ 1.566699 & 1.549193 \end{bmatrix}$$

Covariance matrix of A_{scaled}

$$C = \begin{bmatrix} 1.2 & 1.13265577 \\ 1.13265577 & 1.2 \end{bmatrix}$$

Eigenvalues are 2.33265577 and 0.06734423. Corresponding eigenvectors are $\begin{bmatrix} 0.70710678 \\ 0.70710678 \end{bmatrix}$ and $\begin{bmatrix} 0.70710678 \\ -0.70710678 \end{bmatrix}$

Projecting the scaled data to one dimension using this eigenvector, we get

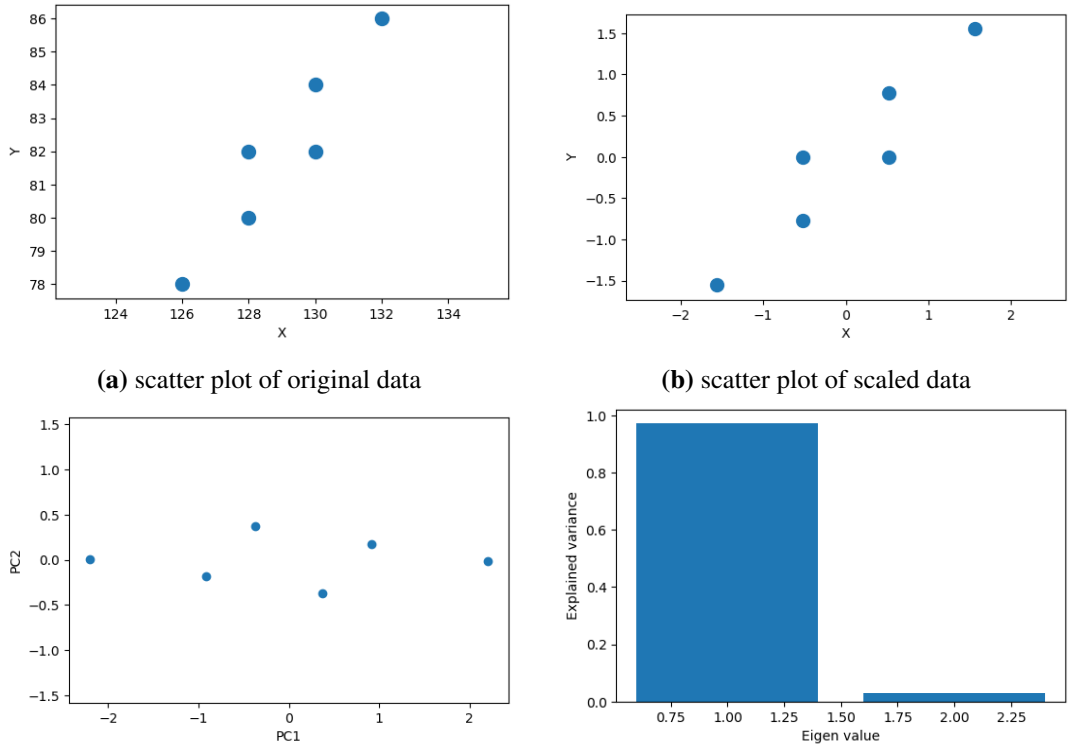
$$Projected_{data} = \begin{bmatrix} -2.20326853 \\ -0.91699703 \\ -0.36927447 \\ 0.36927447 \\ 0.91699703 \\ 2.20326853 \end{bmatrix}$$

The first projected data can be found by the following formula:

(Transpose of eigenvector) * (Feature vector)

for example: $-1.566699 * 0.70710678 + -1.549193 * 0.70710678 = -2.20326835894776$

ann so on.



(a) scatter plot of original data

(b) scatter plot of scaled data

(c) scatter plot of scaled data in PC space

(d) Explained variance of projected data

Fig. 2.3. Step by step visualization of PCA

2.3 Singular Value Decomposition

Singular Value Decomposition (SVD) is one of the dimensionality reduction and denoising techniques that allow an exact representation of any type of matrix[43–45]. It produces an approximate representation with any desired number of dimensions by eliminating the less important parts of the representation. Let A be an m× n matrix of rank r. Then A can be written as the products of three matrices U, ∑, and V. Here U is an m × r column-orthonormal matrix, V is an n × r column-orthonormal matrix and ∑ is a diagonal matrix. The elements of ∑ are called the singular values of A.

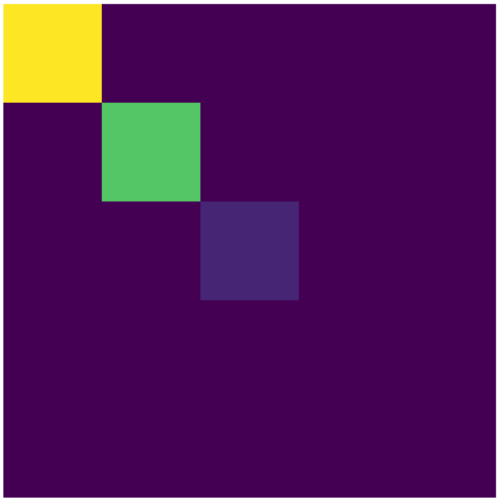
Let us consider the following matrix A. In case that rank $= 1$, the result looks like: $\sigma_1 u_1 v_1^*$,and for rank $= 2$,we decompose the matrix into: $\sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^*$. The image of the matrix A is shown in Fig. 2.4a. Three decomposition matrices U, ∑, and V^* for the matrix A are shown in Fig. 2.4b-2.4d respectively. Also, 1,2,3,4-rank approximation of the matrix A are shown in Fig. 2.5a-2.5d respectively. It is observed that the 3-rank and 4-rank approximations are similar to the original A. By discarding the less important parts

of the representation (two lower singular values), we can reconstruct the original matrix or denoised version of the original matrix. This is the beauty of SVD.

$$A = \begin{bmatrix} 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \end{bmatrix}$$

(a) image of the matrix A

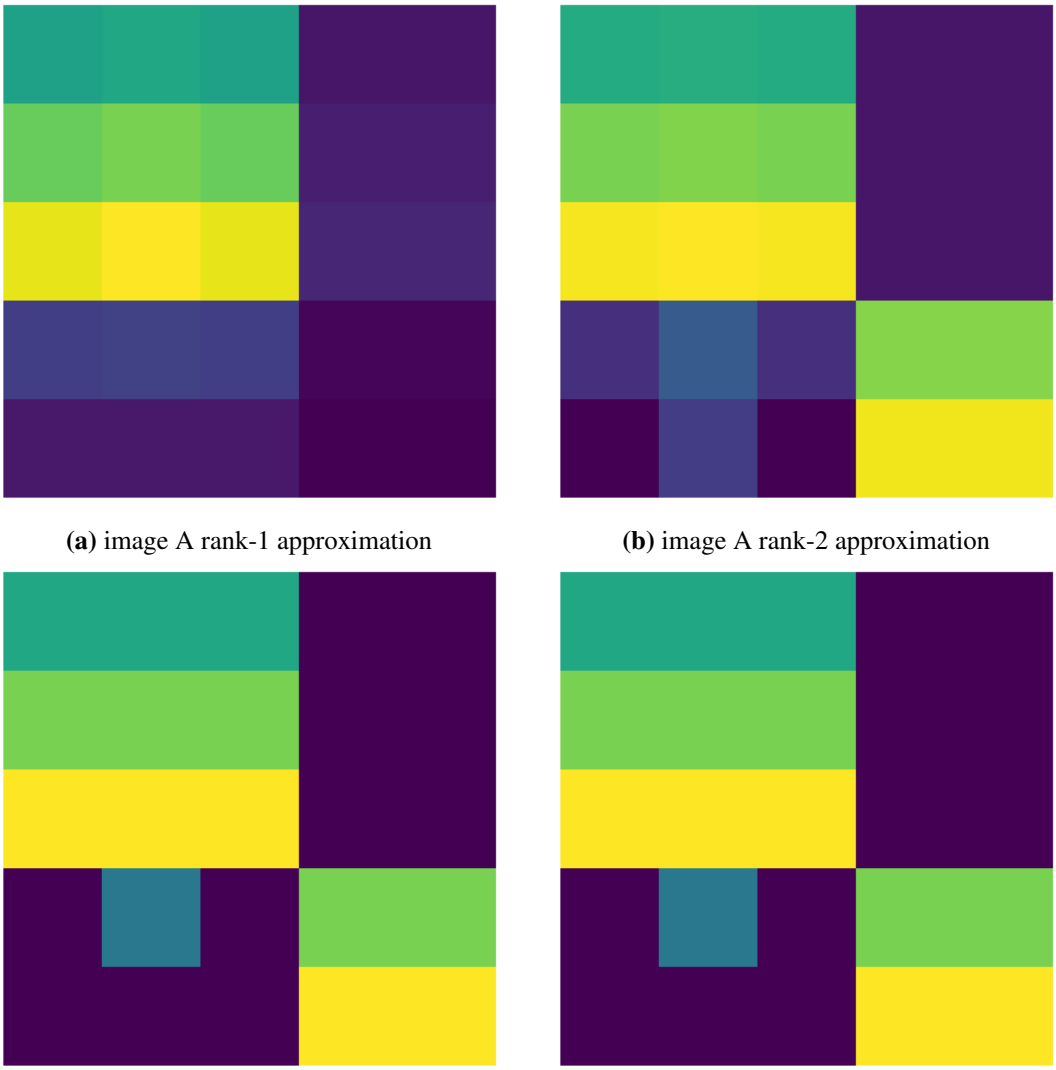
(b) image of the matrix U



(c) image of the matrix $\sum$

(d) image of the matrix V transpose

Fig. 2.4. Step by step visualization of SVD



(a) image A rank-1 approximation

(b) image A rank-2 approximation

(c) image A rank-3 approximation

(d) image A rank-4 approximation

Fig. 2.5. Reconstruction of image by SVD

2.4 Multidimensional scaling (MDS)

Multidimensional scaling (MDS) is a powerful visual representation of data from high dimension to low dimensional space based on the dissimilarities (distances) between sets of instances[46–48]. If the two objects are close together in high-dimensional space, MDS will retain those two objects close together in low-dimensional space. Distance matrix, *PD* between each pair of objects is needed to calculate for MDS. Usually, the Euclidean distance matrix is used, it is the squared value of the distance between objects. But we can use any other distance.

It minimize the following quantity

$$PD(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(d_{ij}^{(X)} - d_{ij}^{(Y)}\right)^2$$

where $d_{ij}^{(X)} = \left\|x_i - x_j\right\|_{L_2}$ and $d_{ij}^{(Y)} = \left\|y_i - y_j\right\|_{L_2}$ are, respectively, the pairwise distances between points i and j in high- and low-dimensional spaces. Here the original high-dimensional data points $X = [x_1, x_2, \ldots, x_n]_{p\times n}, x_i \in \mathscr{R}^p$ map to data points $Y = [y_1, y_2, \ldots, y_n]_{q\times n}, y_i \in \mathscr{R}^q, y << p$ in a low-dimensional space.

2.5 Random Forest

2.5.1 Random Forest regression

A random forest regression(RFR) consists of several decision trees (DT)[49] and the final prediction is the average of all individual trees. RFR is an ensemble machine learning algorithm. It creates several DTs using bootstrapping (random sampling with replacement) from the available data in the training set. Each DT gives its own individual prediction. Averaging all individual predictions gives the RFR prediction shown in Fig. 2.6. So, RFR is better than a single DT algorithm and it enhances the accuracy and decreases the overfitting. A decision tree (DT) is one of the machine learning algorithms that make the tree, based on a set of if-else conditions. DT helps to visualize the data in a better way. DT contains several nodes like parent nodes, child nodes, decision nodes, and leaf nodes. To construct DT, let T be a regression tree that splits at a node t. Suppose s is a proposed split for a variable X that splits t into left and right child/daughter nodes t_L and t_R respectively depending on the following conditions: $X \leq s$ or $X > s$; i.e., $t_L = \{\mathbf{X}_i \in t, X_i \leq s\}$ and $t_R = \{\mathbf{X}_i \in t, X_i > s\}$. Regression node[50] impurity is decided by the sample variance within the node. The impurity of t is defined by

$$\hat{\Delta}(t) = \frac{1}{N}\sum_{\mathbf{X}_i \in t}(Y_i - \bar{Y}_t)^2, \tag{2.1}$$

where $\bar{Y}_t$ is the sample mean for t and N is the sample size of t. Similarly, impurities for the daughter nodes are

$$\hat{\Delta}(t_L) = \frac{1}{N_L} \sum_{i \in t_L} (Y_i - \bar{Y}_{t_L})^2, \tag{2.2}$$

$$\hat{\Delta}(t_R) = \frac{1}{N_R} \sum_{i \in t_R} (Y_i - \bar{Y}_{t_R})^2, \tag{2.3}$$

where $\bar{Y}_{t_L}$ is the sample mean for t_L and N_L is the sample size of t_L and analogous for right daughter node t_R. The decrease in impurity under the split s for X equals

$$\hat{\Delta}(s,t) = \hat{\Delta}(t) - \left[\hat{p}(t_L)\hat{\Delta}(t_L) + \hat{p}(t_R)\hat{\Delta}(t_R)\right], \tag{2.4}$$

where $\hat{p}(t_L) = N_L/N$ and $\hat{p}(t_R) = N_R/N$ are the proportions of observations in t_L and t_R, respectively. Maximizing the $\hat{\Delta}(s,t)$ to find the best split-point s,which is equivalent to minimizing the quantity $\left[\hat{p}(t_L)\hat{\Delta}(t_L) + \hat{p}(t_R)\hat{\Delta}(t_R)\right]$. The splitting process for a tree is considered finished when the number of training instances at a node falls below or equal to a specified minimum threshold, say n. At this point, the node becomes a leaf node, and the prediction made at that leaf node is typically the average of the target variable values (Y values) of the training instances that reach that specific leaf node (Fig. 2.7).

For example, consider X = [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15] and Y = [1,1.2,1.4,1.1,1,5.5,6.1,6.7,6.4,6,6,3,3.2,3.1,3]. Here, Y is our target. We construct a DT based on X values. We split the X based on the sample variance. The spliting lines at X = 5.5, 3.5, 11.5 and so on respectively. Scatter plots with spliting lines are shown in the Fig. 2.7. Corresponding DT is shown in the Fig. 2.8. In Fig. 2.8, mse is the mean square error which is defined in the Eqn. 2.1 and value is the average of the corresponding Y values at that node.

Furthermore, I explain the procedure of random forest regression by using a simple illustrative example in Fig. 2.9 as follows: we have measured a continuous variable y as a function of an observable x. Our goal is to describe the relation $y = f(x)$ where f is the function we aim to approximate, with a statistical model. There are couples of strategies

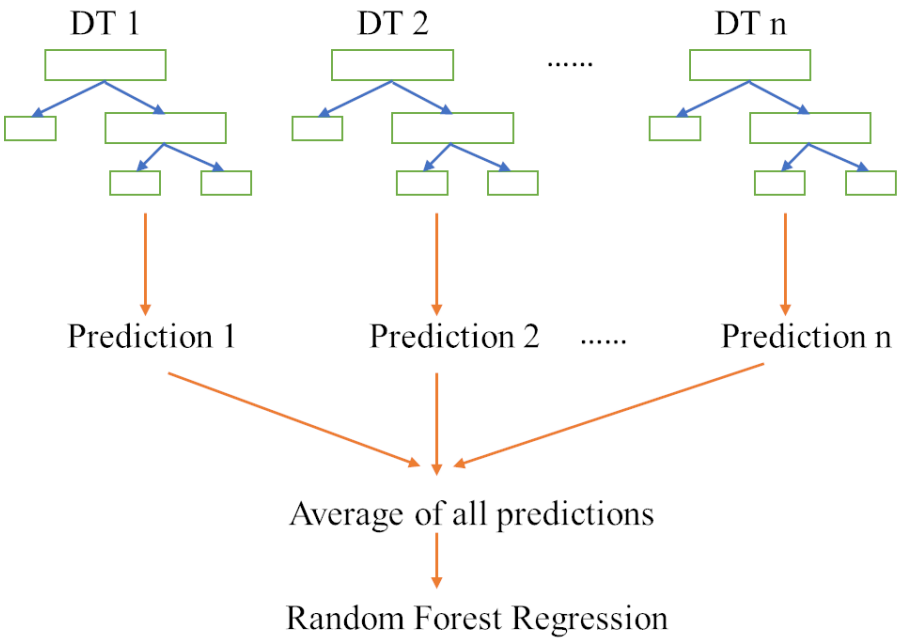


Fig. 2.6. Schematic diagram of Random Forest Regression.

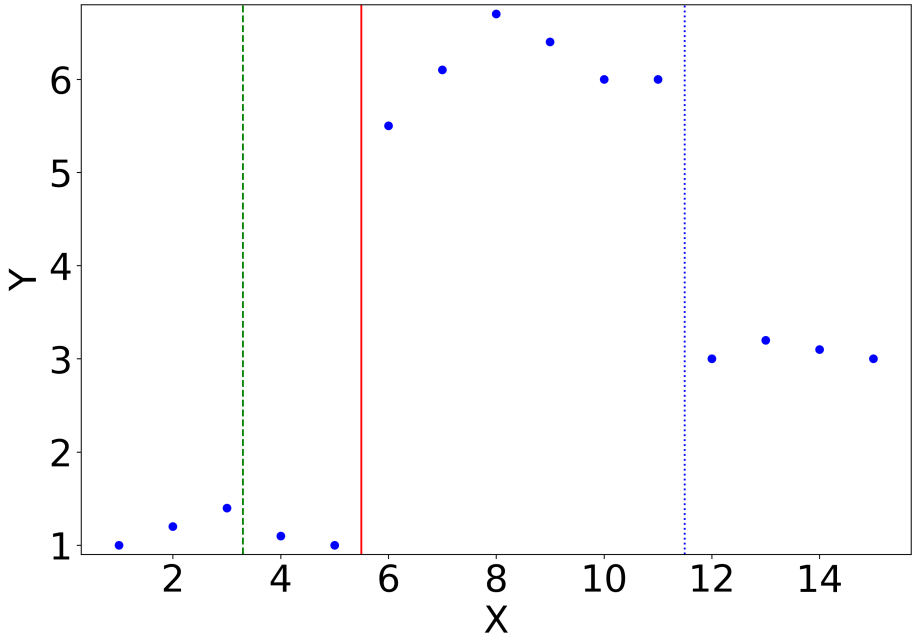


Fig. 2.7. Scatter plots with spliting lines.

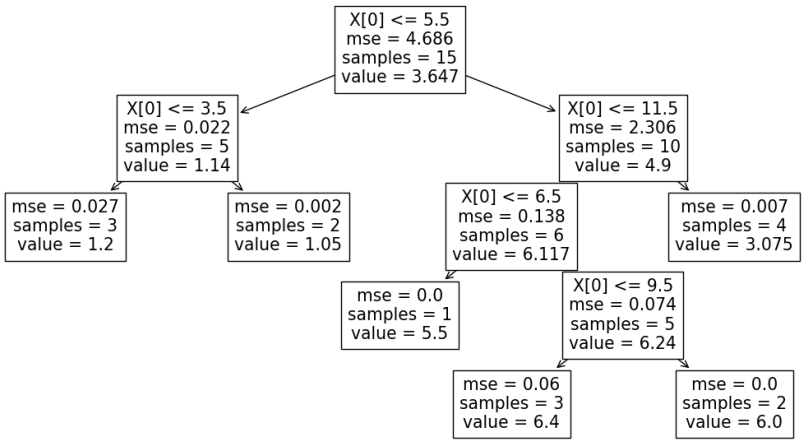


Fig. 2.8. Decision tree.

that can be employed to approximate this function. If we have some prior knowledge about the form of f, for instance if we know it is linear or polynomial in nature, a parametric model such as linear regression or polynomial regression can be a good fit. On the other hand, if no prior knowledge on f is available, we can employ non-parametric regression model such as random forest that does not make assumptions on the functional form of f. To understand random *forest*, we need to introduce its fundamental building block — decision *tree*. Decision trees (DTs) create a model in the shape of a upside-down tree with a set of connected nodes. This hierarchical structure begins with the root node at the top of the tree, which holds the initial data set. From this root node, the data is partitioned based on a chosen value of x, splitting the data into two distinct child nodes: a left child node and a right child node. To contextualize this process, Fig. 2.9A, we show the approximated function mapping x to y, estimated by a tree containing one root node and two child nodes, essentially a tree formed by one data split guided by the following condition : $x \leq i$ with $i \in [\min(x), \max(x)]$. If the condition is fulfilled, the data reach the left child node; otherwise, they go into the right child node. The estimated function, denoted as $\hat{y} = \hat{f}(x)$, looks like a Heaviside step-wise function. This is because the predicted values $\hat{y}$ provided by a tree are the mean values of y for those x values that falls within a particular terminal nodes, either left or right. In simpler terms, given a tree with two terminal nodes, for each x

value to be predicted, the assigned prediction $\hat{y}$ will either be the average value of y for the subset of observations that reach the left terminal node or the right terminal node, depending on whether the x value satisfies the splitting condition at the root node. We can also mention that if a tree is only composed of the root node, i.e. a unique node, the approximated function $\hat{f}$ is a constant function whose constant value is the mean value of y. During the training phase of a decision tree (DT), the goal is to find the best splitting conditions of the data set that minimize the root mean squared error defined as (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, with the number of observations n. Fig. 2.10, for example, displays the RMSE as a function of the x values splitting condition, for the tree with one root node and two terminal nodes. To get a more accurate approximation of the function f than a Heaviside step-wise function, the tree is growing up with multiple data splits until a stopping condition is met. Commonly, a stopping condition can specify that a node will not be split further if it contains fewer observation than a specified minimum criterion or if a set of maximum tree depth is reached. Typically, the minimum sample size at terminal nodes of the tree, and the minimum number sample of split, are the two main hyper-parameters that control the quality of the approximated function, $\hat{f}$. Fig. 2.9A, the function becomes more precise for 5 or 15 data splits, compared to just one split. A generalization of the decision *tree* regression is random forest regression (RFR) that constructs an ensemble of DTs. The term "random" in RFR refers to the method of bootstrap sampling with replacement [50–52]. In RFR, each DT is constructed with a boostrapped sample drawn from the original data set. This method ensures that the collection of trees presents a certain level of diversity which is known to be a "safeguard" against overfitting the training data. The final approximated function made by RFR, $\hat{f}_{\mathrm{RF}}$ is an aggregation of the estimated functions given by individual DTs such that $\hat{f}_{\mathrm{RF}}(x) = \frac{1}{q}\sum_{i=1}^{q} f_q(x)$ with the approximated function of a single tree $f_q(x)$ and the total number of trees in the forest q. Fig. 2.9B shows, for example, the estimation of the function f, given by random forest and a single decision tree. The set of DTs (random forest) can approximate the underlying function or trend free from overfitting without choosing a parameter such as order in polynomial regression. In this research, I employed default hyperparameters values of RFR as follows: The number

of DTs in the random forest q is 100, the minimum number of samples belonging to a leaf node is 1.

2.5.2 Random Forest classification

For the classification problem, Gini impurity is used instead of variance. For the random forest binary classification, the optimal split is evaluated at each node τ within the binary trees T , using the Gini impurity $i(\tau)$ which is defined as[53]

$$i(\tau) = 1 - p_1^2 - p_0^2$$

where $p_k = \frac{n_k}{n}$ is the fraction of the n_k samples from class $k = \{0,1\}$ out of the total n samples at node τ. Its decrease Δi that results from splitting the node into two sub-nodes τ_l and τ_r by setting a threshold t_θ on variable θ is defined as

$$\Delta i(\tau) = i(\tau) - p_l i(\tau_l) - p_r i(\tau_r)$$

The decrease in Gini impurity resulting from this optimal split $\Delta \mathrm{i}_\theta(\tau, T)$ is then recorded for the best split. For the classification case, the final prediction is based on the majority votes of all individual trees.

In conclusion, this chapter provides an overview and discussion of various useful algorithms that will assist as the foundation for the evaluation conducted in later chapters. These algorithms have been carefully chosen based on their relevance and applicability to the research objectives.

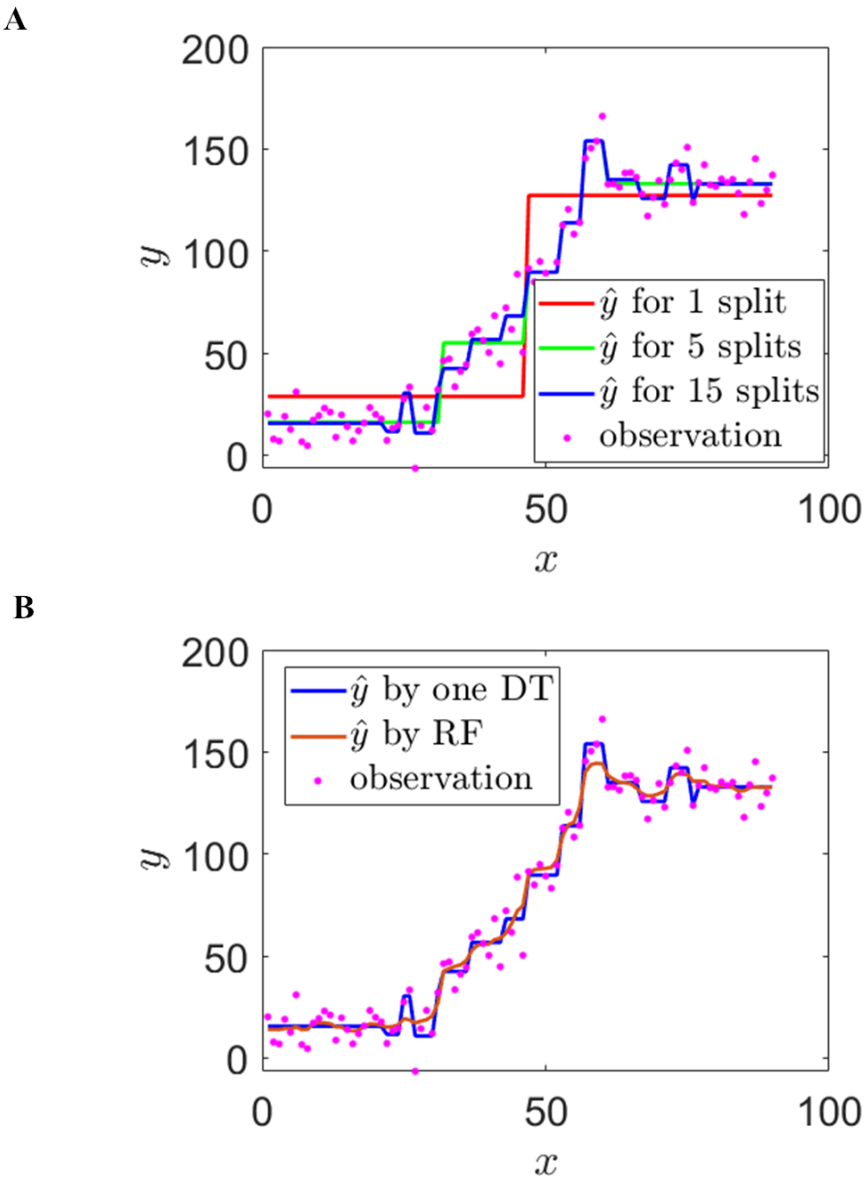


Fig. 2.9. (A) A scatter plot of "observed" data with the approximate regressions by a DT in terms of 3 different splits (1, 5, and 15). (B) A scatter plot with the approximate regressions by a single DT with 15 splits and by a set of 100 DTs.

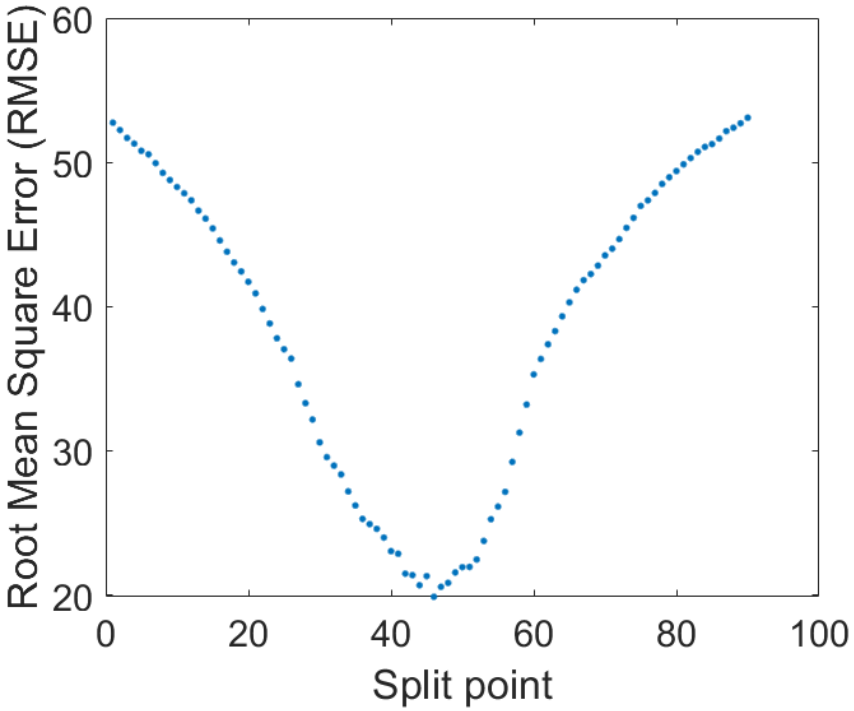


Fig. 2.10. The root mean squared error (RMSE) as a function of the location at which the point to split the given data set is determined for the data set used in Fig. 2.9.

3

Raman data preprocessing

3.1 Data preprocessing

Data preprocessing in Machine Learning is the most important and tedious step that aids to enhance the quality of data. It is a technique of cleaning and organizing the raw data to make it reasonable for building and training Machine Learning Algorithms. The measured Raman spectra are corrupted by several phenomena like fluorescence background, cosmic spikes and white noise, etc.[1],[54] shown in Fig. 3.1. Before analysis it is very important to clean Raman data from all types of corruptions for further analysis.

Fig. 3.2 is displayed a schematic diagram of the three preprocessing workflows for Raman image analysis. While wavenumber calibration is an important step in the preprocessing of Raman data, it is mentioned separately as we observed a shift in the peak

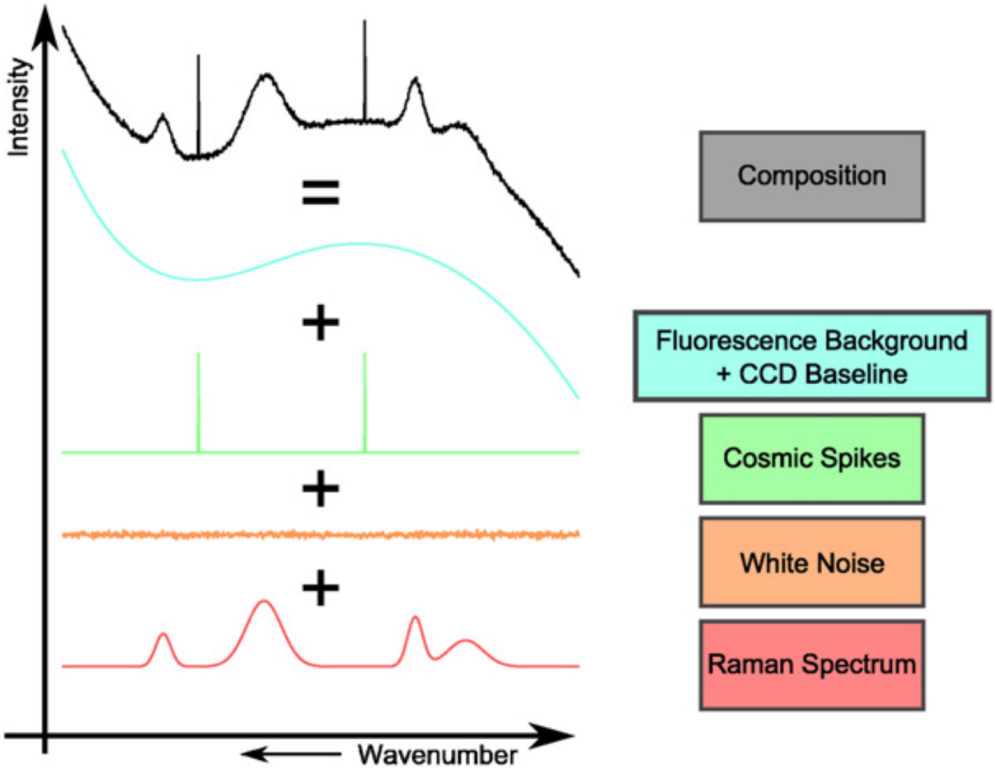


Fig. 3.1. Spectrum composition[1].

position of spectra obtained from different cells located at different spatial positions when applying wavenumber calibration independently of the position along the illumination axis ζ (referred to here as 'without wavenumber calibration' or 'uncalibrated')(Fig. 3.3). As a consequence some non-negligible spatial dependence appears in both some of the resulting Raman images and PC score images when the standard preprocessing without wavenumber calibration is applied. Then I fixed our standard preprocessing workflow along with the position-dependent wavenumber calibration along the illumination axis ζ to minimize spatial dependencies between spatial axis and chemical intensity distribution in the following analysis. Furthermore, I present a detrending scheme based on random forest regression to enhance the differentiability of the Raman signals between FTC-133 and Nthy-ori 3-1.

3.1.1 Wavenumber calibration along illumination line

In line-illumination microscope experiments, the wavenumber axis is calibrated based on the reference sample spectrum and for each individual pixel along the line direction. This accounts for potential drifts in Raman peak positions that may occur along the line axis.

The calibration protocol assumes a consistency of Raman shift drift errors in the scanning direction and is established with a single Raman line measurement of ethanol, using a 0.5-second exposure time. Consequently, 400 spectra of ethanol are recorded at the CCD detector and organized into a matrix of dimensions $(m, \nu) = (400, 910)$, with 910 being the number of pixels along the wavenumber axis. For each of the 400 ethanol spectra, the seven theoretical Raman peaks of ethanol solution (884, 1052, 1096, 1454, 2880, 2930, 2974 cm^{-1}) (Fig. 3.4) were detected at different pixel indices, and a third-order polynomial model was used to estimate the continuous nonlinear relationship between pixel indices and Raman shifts. Each estimated third order polynomial model provides a new wavenumber axis of size 910 pixels resulting in the estimation of 400 new wavenumber axis in total. A cubic spline model denoted by f_{kl}, with the spatial position along the scanning axis $k \in [1, 240]$ and the spatial position along the illumination line axis $l \in [1, 400]$, learns the mapping between the new Raman shift axis ν_j to the Raman intensity of individual spectrum of a Raman image as $f_{kl}\left(\nu_j\right)$. These different cubic spline models are then used to interpolate all the individual spectrum of a Raman image on a consistent linear grid of spectral resolution 3.8 cm^{-1} to have a common wavenumber axis between all spectra of a Raman image and between different measurements.

3.1.2 Raman data preprocessing steps

Prior to analysis, Raman images underwent preprocessing with a standard protocol aimed at minimizing known spectral artifacts. The preprocessing workflow consisted of several steps: (1) cosmic ray removal: cosmic rays appear as intense spike in Raman spectra at random position. Their localization can be expressed as an outlier detection problem, where in each 2D Raman image u at each wavenumber ν_i, a pixel is considered as corresponding to a cosmic ray if its intensity exceeds a threshold of $\mu(u_{\nu_i}) + 8\sigma(u_{\nu_i})$, where $\mu(u_{\nu_i})$ is the mean intensity of the Raman image u at ν_i and $\sigma(u_{\nu_i})$ is its standard deviation. Cosmic ray intensity is then replaced by the mean intensity of the 9 closest neighboring pixels, including the cosmic ray's value. This cosmic ray detection and replacement is performed recursively for each wavenumber until no more cosmic rays are detected. (2) Bias cor-

rection: Constant value 520 photon counts was subtracted from each intensity due to the intrinsic bias of our device. Then the position-dependent wavenumber calibration along illumination axis was performed, explained in the previous section. (3) Noise reduction: the Raman spectra can be degraded by various types of noise such as read-out noise, fluorescence background noise, Raman photon noise, and dark current noise. To improve the signal-to-noise ratio of Raman images, singular value decomposition (SVD) denoising is used by keeping the first 8 singular value components [44]. Intensity distributions of four well known peaks are shown in Fig. 3.5. (4) Fluorescence background correction: Raman spectra are distorted by a baseline fluorescence background originates from the substrate, autofluorecence molecules in samples or other elements. We reduced this fluorescent baseline in each individual Raman spectra of a Raman image by using the modified polynomial algorithm (modpoly) [55] of order 8. Average spectrum of 6 cells are shown in Fig. 3.6 and Fig. 3.7 after denoising and baseline corrections respectively. (5) Data normalization: Raman spectra are subject to diverse multiplicative effects such as the varying number of molecules at different positions, laser power fluctuations and focus drift among others which modify Raman intensity. To make spectra comparable from experiments to experiments or positions to positions total intensity normalization was employed. The normalization procedure involves dividing the intensity $u(k,l,\nu_i)$ of each individual spectrum in the Raman image by the constant $\sum_{\nu_i=1}^{\nu} u(k,l,\nu_i)$. Normalization was performed over the wavenumber range 581 cm^{-1} to 3025 cm^{-1} after truncation of the silent regions in the range (1880-2805 cm^{-1}). See also in Appendices Parameters selection for SVD and baseline correction section.

Note that there are some spatial correlation between each principal component and spatial axes ζ and ξ after standard preprocessing shown in Fig. 3.8 and Fig. 3.9. My target was to remove this dependency by detrending each principal component.

3.1.3 Proposed detrending scheme

After standard preprocessing with position-dependent wavenumber calibration, the unfolded (=preprocessed) Raman image of size (nm,ν) is expanded in an orthonormal ba-

sis known as Karhunen-Loève (K-L) basis or principal component (PC) basis [39, 40] to translate the Raman image as a set of 2D maps of PC scores of $n \times m$ pixels, denoted as $Q_i(k,l)$ with $1 \leq k \leq n$ and $1 \leq l \leq m$, or simply Q_i otherwise noted. These maps reflect a series of spectral variations over the physical space buried in the Raman image, which can reflect the presence of slowly varying change of intensity related to non-homogenous illumination effects as illustrated by the example of the 5th PC score Q_5 (Fig. 3.10) with a gradient of intensity. To further visualize these effects, we plot individual $Q_5(k,l)$ as a function of the scanning axis ξ (corresponding to k), and the laser illumination line axis ζ (corresponding to l), respectively (See Fig. 3.11C and Fig. 3.11E).

We suppose that, for Raman images without spatial degradation, the individual PC score Q_i should be non-correlated to both of illumination and scanning axes, leading to symmetric distributions centered around zero between each PC score and these spatial axes. However, some correlation between the first tens ($\ll \nu$) principal component scores and the spatial axes remain even after the application of position-dependent wavenumber calibration (Fig. 3.11A) (Note again that, without calibration, artificial spatial correlations are much more significant, e.g., Fig. 3.8). Importantly in the PC orthonormal basis, the PC score are mutually uncorrelated, as shown by the correlation matrix of the full set of PC scores (Fig. 3.13B).

This implies that the application of a nonlinear detrending correction is straightforwards, i.e., the detrending operation to Q_i does not affect to one another. This is not true if we correct the individual Raman shifts as the spectral features of a Raman image are mutually correlated among them, as seen by the correlation matrix of the preprocessed Raman image (Fig. 3.13A).

The workflows to detrend the spatial correlation Q_i along each spatial axis is as follows: we first employ a series of random forest regression (RFR) models [50, 52] to estimate the slowly varying relation between each Q_i and the illumination axis ζ. Here we chose the random forest regression model to estimate the underlying trend because of its nonparametric nature, which is more adaptable in estimating unknown nonlinear relations compared to parametric models such as polynomial regression. An example of the esti-

mated trend by RFR is given for the 5th PC score Q_5 along ζ (Fig. 3.11C). The spatially averaged trend in Q_i along the illumination axis ζ (denoted by $\hat{Q}_i(l)$) that could not be removed by position-dependent wavenumber calibration are then subtracted from each Q_i. That is, $Q'_i(k,l) = Q_i(k,l) - \hat{Q}_i(l)$ for all pairs of k and l. Afterwards, the same process is repeated to remove the spatial correlation along the scanning axis ξ. That is, a new series of RFR are performed to estimate the correlation to the scanning axis ξ (denoted by $\bar{Q}_i(k)$) remaining in Q'_i(Fig. 3.11E). The final correction of Q_i (denoted by $\bar{Q}_i$) is then given by $\bar{Q}_i(k,l) = Q'_i(k,l) - \bar{Q}_i(k)$ for all the pairs (k,l). However, for the PC 100th,there is no improvement after detrending because PC100 is mainly noise shown in Fig. 3.12. The detrended 2D PC maps along both the illumination and scanning axes on the top of position-dependent wavenumber calibration are then translated to a detrended 3D Raman image with size (n,m,ν). The advantages of choosing random forest regression (RFR) over the averaged PC score, polynomial fitting of different orders are presented in Appendices (Advantages of the random forest model with other models section). My proposed scheme was performed on a Raman image of Dimethyl sulfoxide (DMSO) (see Appendices in Measurement of Dimethyl sulfoxide (DMSO) section).

Average spectrum of 6 cells based on three schemes are shown in Fig. 3.14. One can easily see the improvement of my detreding scheme compared to other two schemes. My scheme reduces the variations of the intensity of spectra at each Raman shift in a Raman image. This is reasonable because 6 cells are coming from one Raman image and so their spectral difference should be low.

In conclusion, this chapter provides a detailed clarification of the proprosed developed detrending scheme for Raman data preprocessing, highlighting its efficacy in improving spectral quality. The comparison among the proposed scheme and conventional approaches demonstrates a notably reduction in spectral differences, particularly for spectra with the same phenotype.

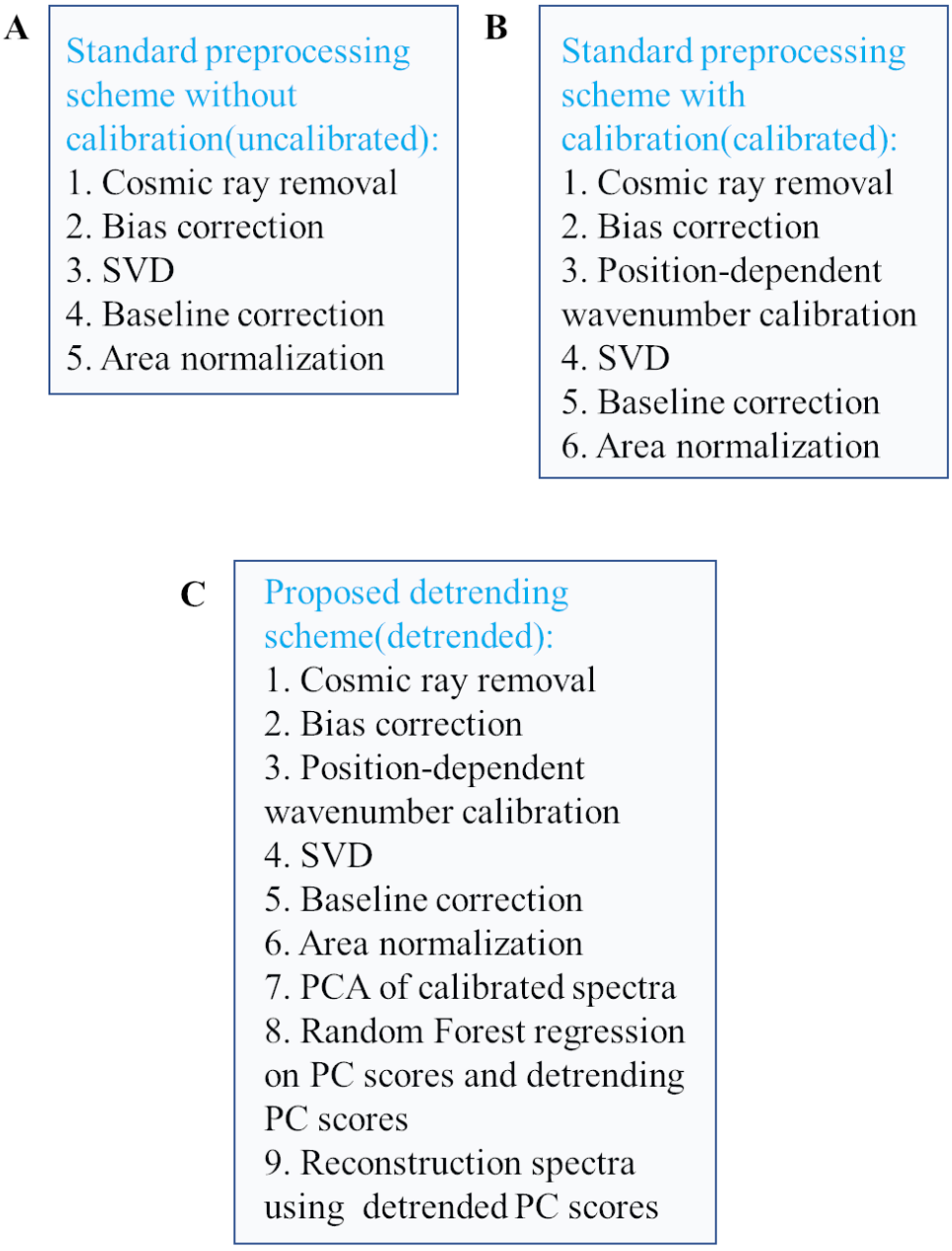


Fig. 3.2. Three preprocessing work flows

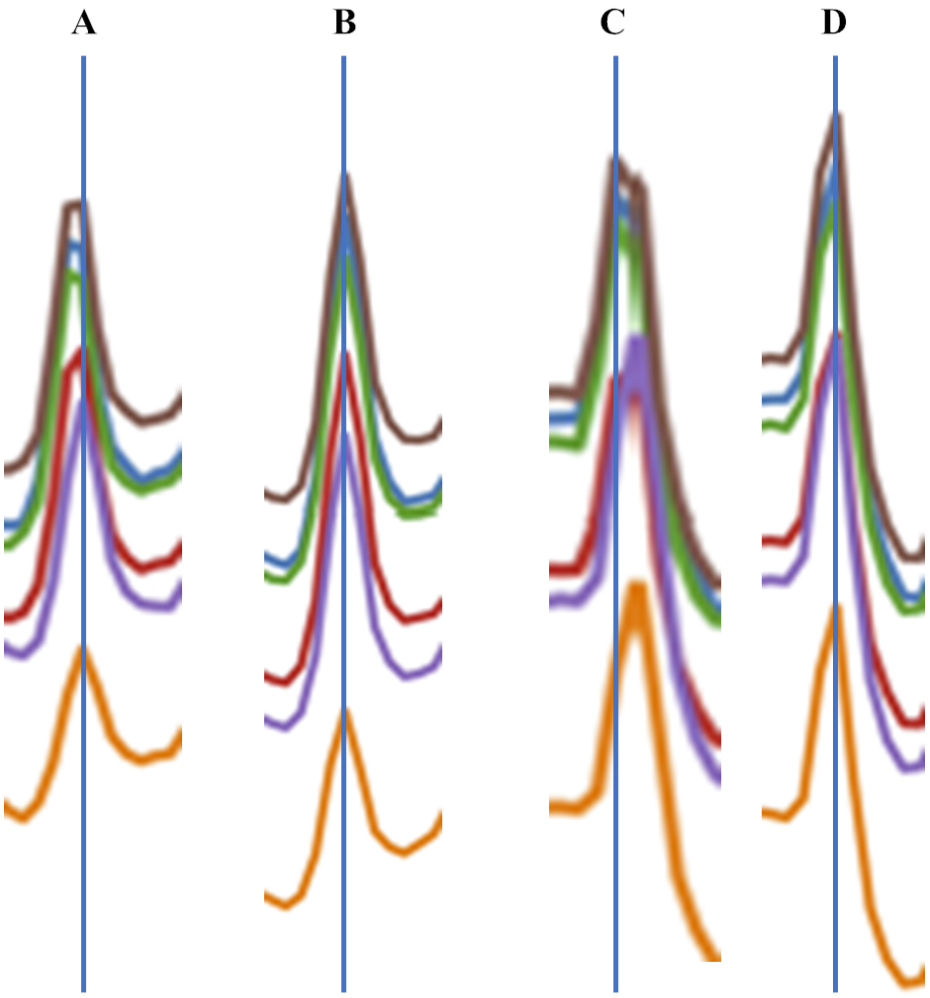


Fig. 3.3. Enlarged peak positions of the averaged spectra of six cells contained in a Raman image of FTC-133(#2) at two Raman shifts (A and C) without wavenumber calibration, (B and D) with the position-dependent wavenumber calibration.

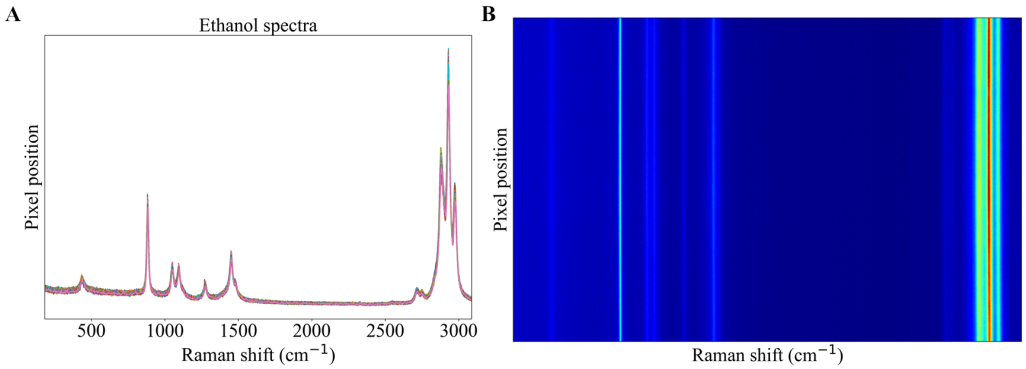


Fig. 3.4. (A) Ethanol all spectra (B) image plot of peak positions for ethanol spectra.

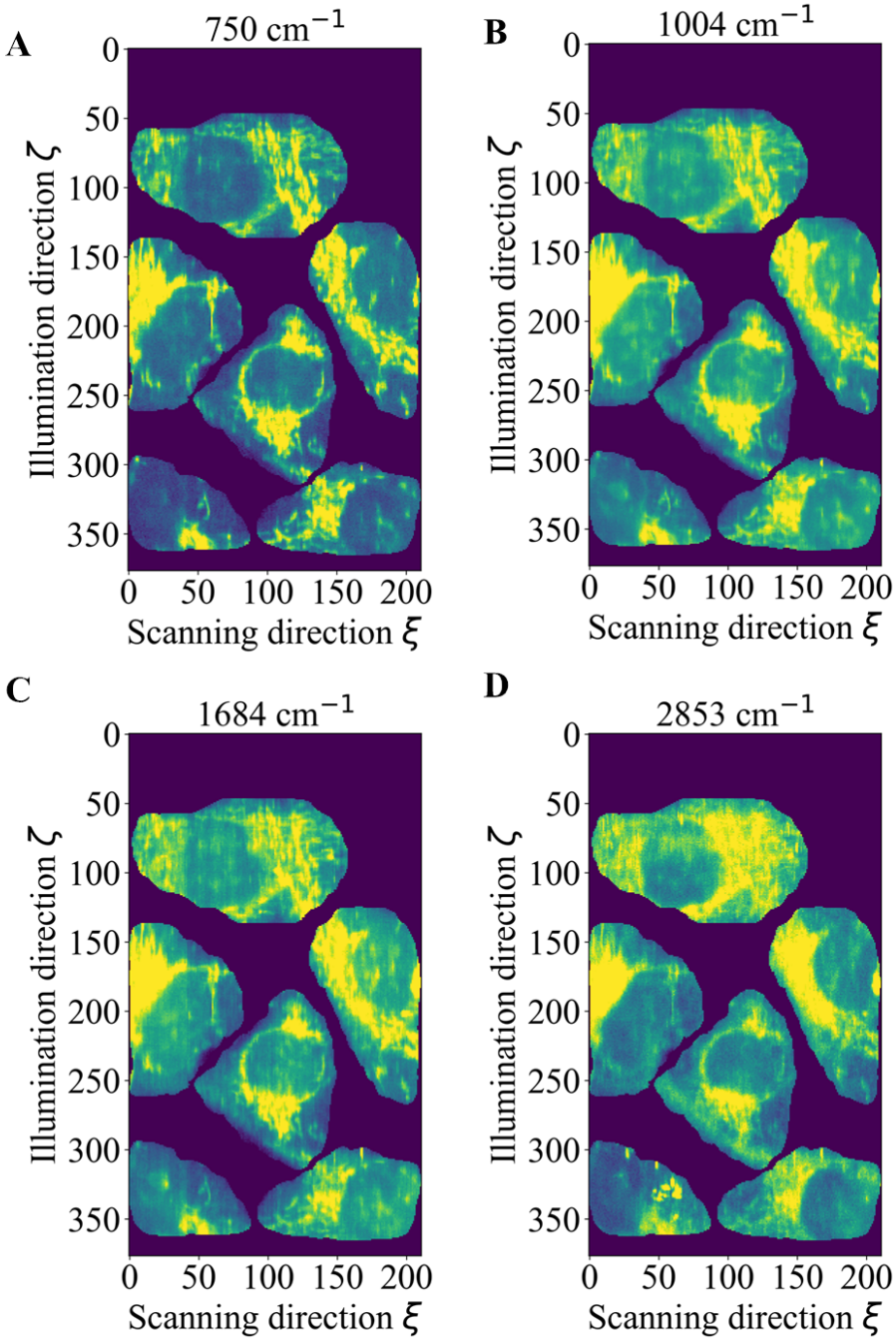


Fig. 3.5. Intensity distribution in a space domain at four peaks after denoising without wavenumber calibration: (A) cytochrome, (B) phenylalanine, (C) protein,and (D) lipid.

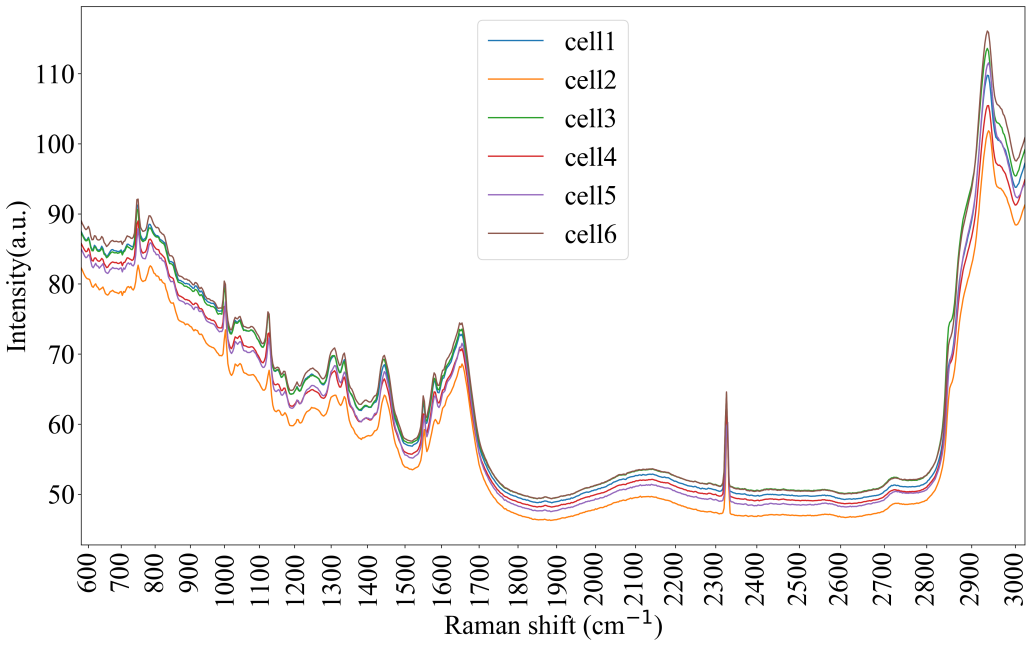


Fig. 3.6. Average spectrum of 6 cells after denoising.

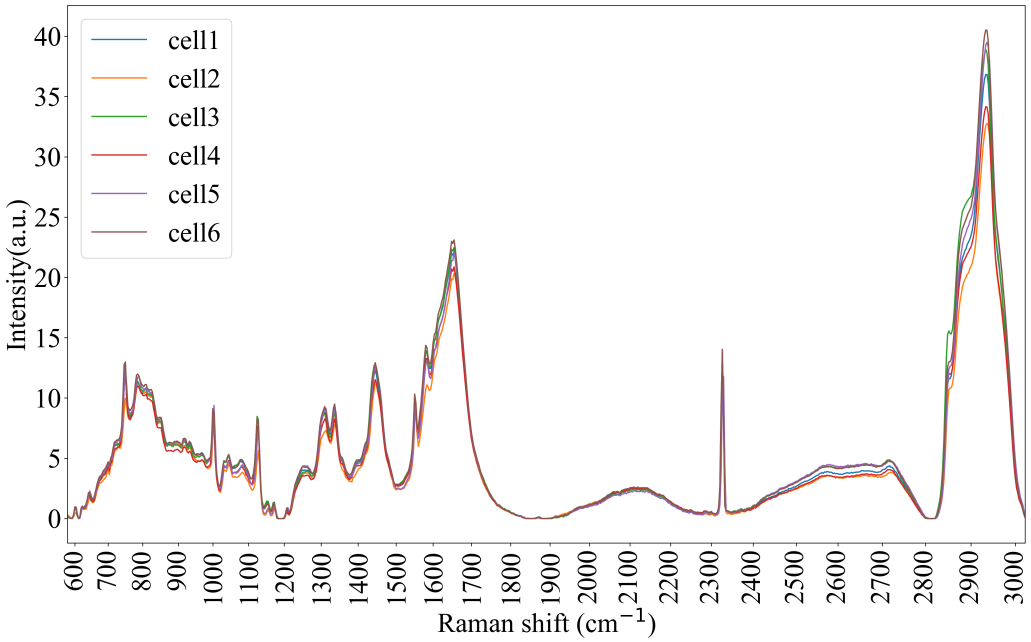


Fig. 3.7. Average spectrum of 6 cells after denoising.

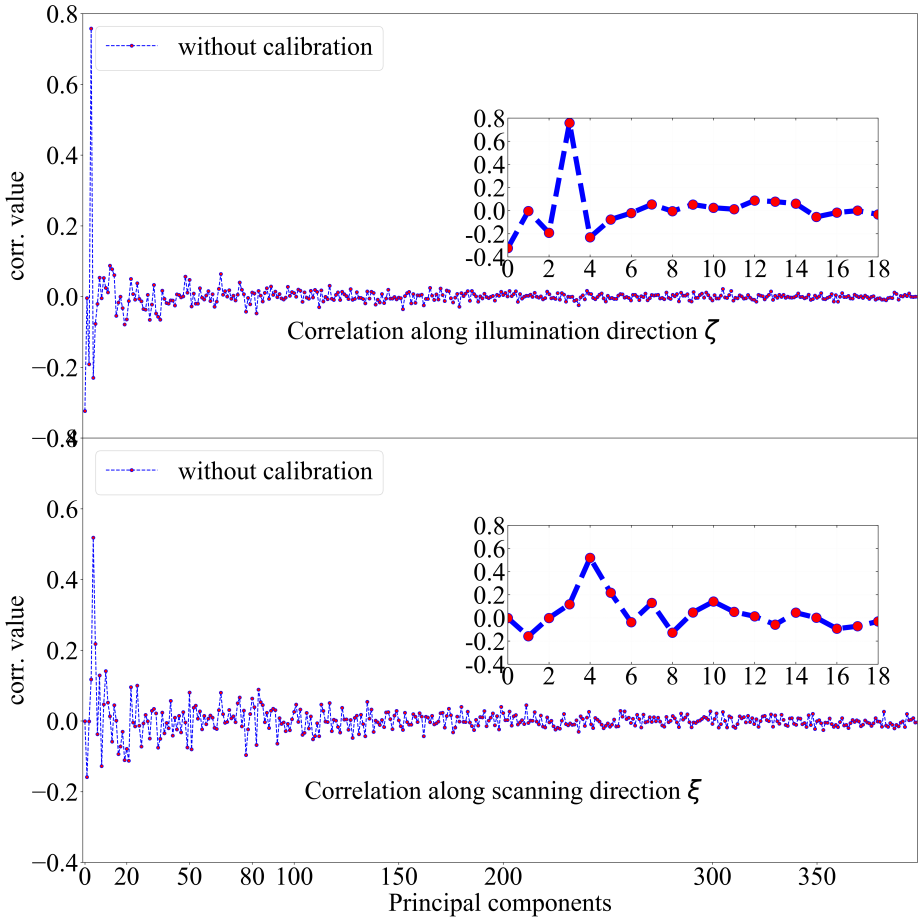


Fig. 3.8. The Pearson correlation coefficients between the spatial coordinates, illumination and scanning axes, and the images of PCs of a Raman image of FTC-133(#2) preprocessed by standard preprocessing without wavenumber calibration.

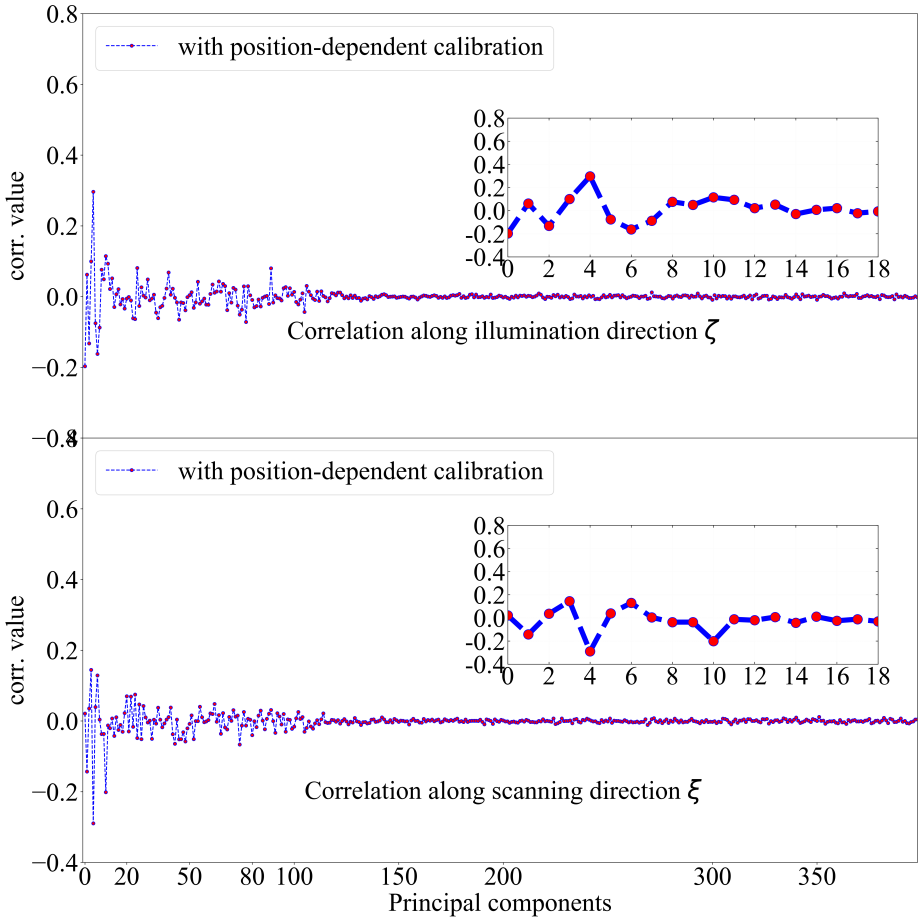


Fig. 3.9. The Pearson correlation coefficients between the spatial coordinates, illumination and scanning axes, and the images of PCs a Raman image of FTC-133(#2) preprocessed by standard preprocessing with the position-dependent wavenumber calibration.

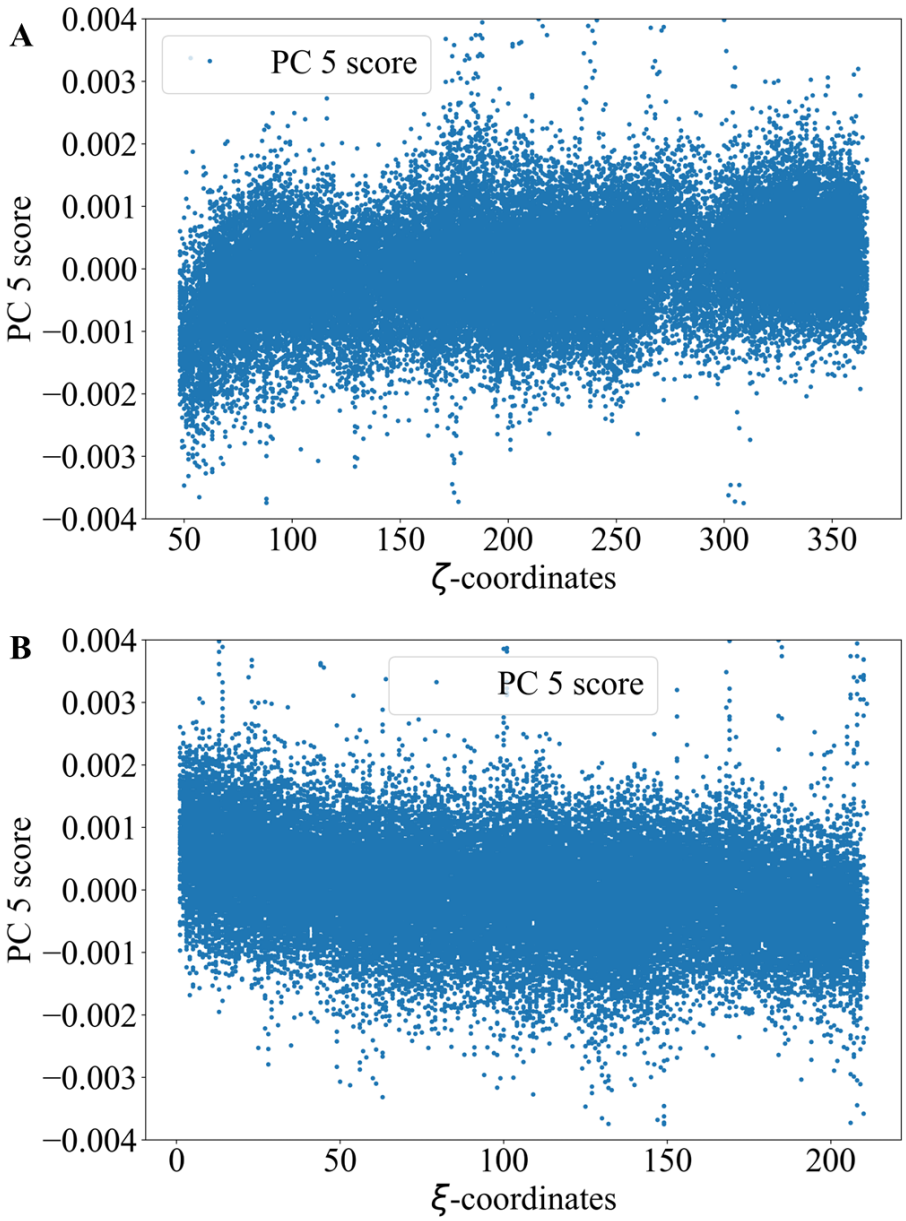


Fig. 3.10. (A) Scatter plot of PC 5 score and ζ-coordinates, (B) Scatter plot of PC 5 score and ξ-coordinates.

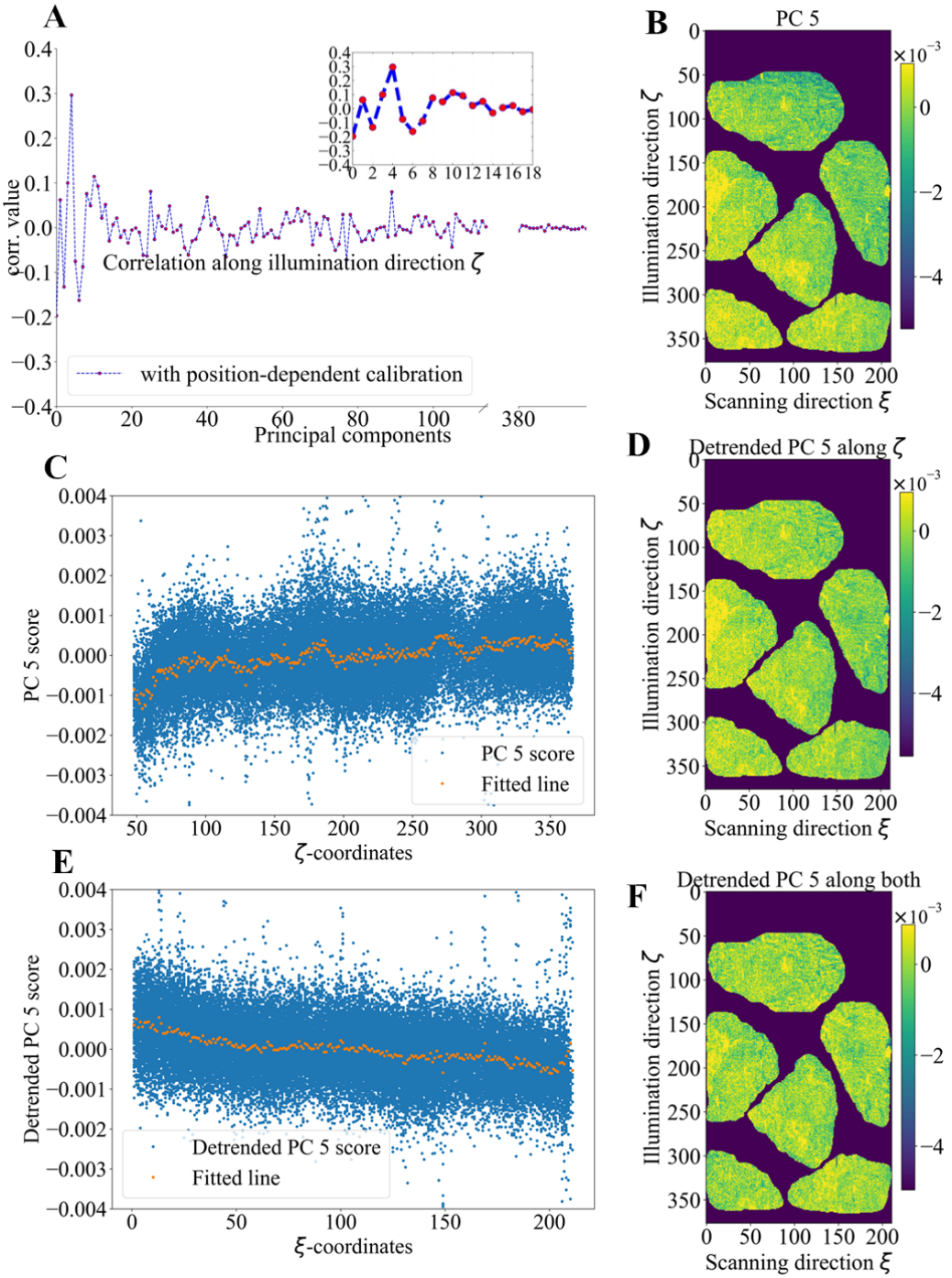


Fig. 3.11. (A) Correlation between illumination axis coordinates and PCs (B) PC 5 scores value distribution in a space domain after standard preprocessing with position-dependent calibration (C) Scatter plot of PC 5 score and ζ-coordinates with RF regression line (D) detrended PC 5 scores value distribution along ζ axis correction in a space domain (E) Scatter plot of detrended PC 5 score and ξ-coordinates with RF regression line (F) detrended PC 5 scores value distribution along both axes correction in a space domain.

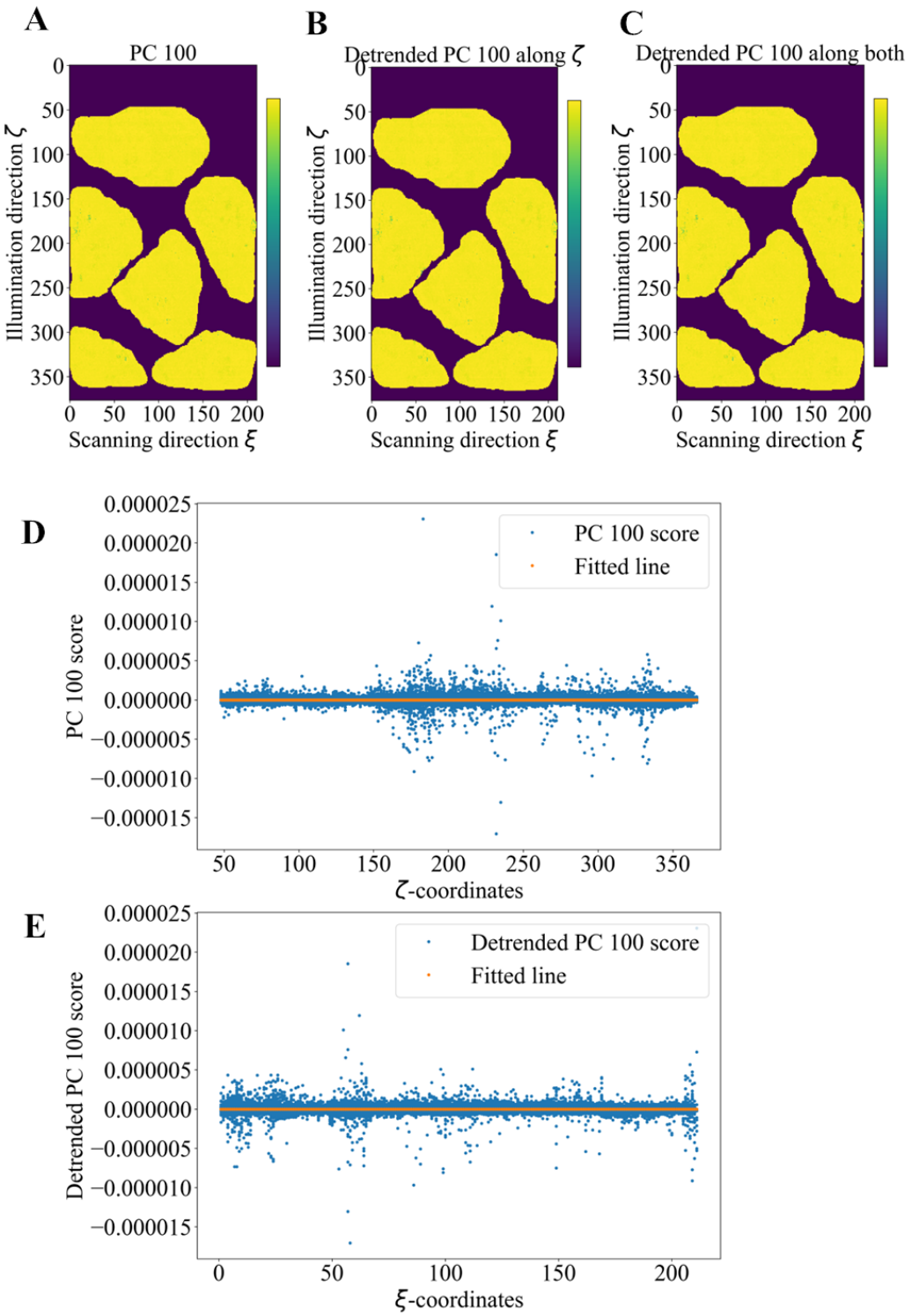


Fig. 3.12. (A) PC 100 scores value distribution in a space domain after standard preprocessing with position-dependent calibration (B) detrended PC 100 scores value distribution along ζ axis correction in a space domain (C) detrended PC 100 scores value distribution along both axes correction in a space domain(D) Scatter plot of PC 100 score and ζ-coordinates with RF regression line (E) Scatter plot of detrended PC 100 score and ξ-coordinates with RF regression line.

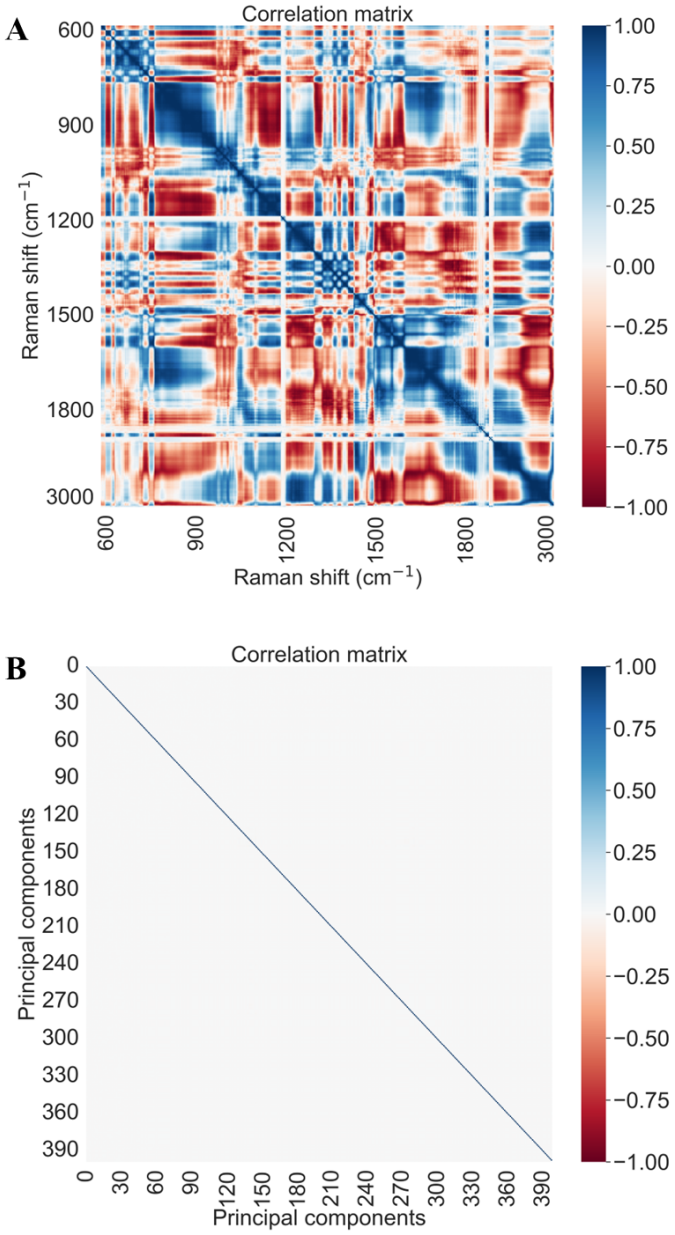


Fig. 3.13. Average spectrum of 6 cells after normalization.

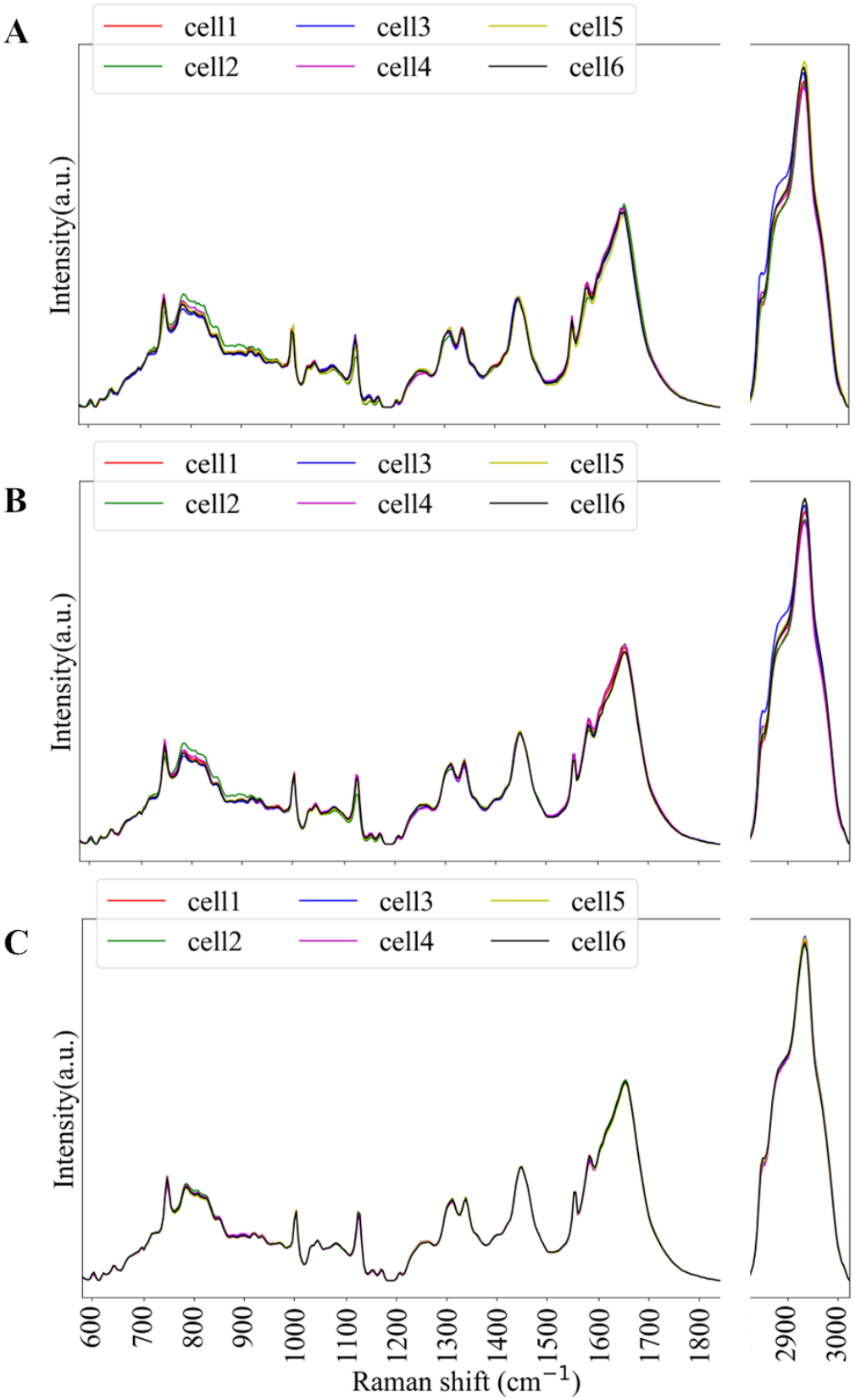


Fig. 3.14. Average spectrum of 6 cells:(A) without wavenumber calibration. (B) the position-dependent wavenumber calibration. (C) the detrending scheme.

4

Differentiability of cell types enhanced by detrending non-homogeneous pattern in line-illumination Raman microscope

Raman microscopy is a label-free, vibrational imaging technique that reflects the underlying, unique spectral features of molecules constituting a sample to measure [56–58]. Despite its potential for use in areas such as disease diagnosis [59, 60], treatment monitoring [61], drug design [62], and cell therapy development [63], the practical application of Raman spectroscopy in clinical settings faces challenges with regards to the relatively long acquisition time of Raman images [64–66] and the lack of data standardization [67] protocols between different microscope systems and experiments.

The former is the consequence of the weak nature of Raman scattering, which requires a long exposure time to capture enough signal for analysis. The latter is influenced by various instrumental and experimental factors such as optics, sample preparation, laser power fluctuations, spectrometer drifts, autofluorescence, and multiple sources of noise. Multi-step preprocessing workflows have been proposed to remove such artifacts in raw Raman data. The workflows prioritize an objective design that involves optimizing cost functions or quality parameters to assess the effectiveness of the preprocessing [1, 68, 69].The traditional preprocessing techniques for Raman tabular data such as cosmic ray removal, spectrometer calibration, denoising, baseline correction, and normalization have proven to be effective in reducing setup dependencies and improving data comparability[70–72].

However, the standard preprocessing workflows that considered only the spectral dimension in the process may lead to only suboptimal correction and the overlook of important artifacts present in the spatial dimension of Raman images such as non-uniform illumination, focus drift, and stripes. Here, non-uniform illumination refers to spatial variation in intensity of the laser source that is used to scan a sample. Line-illumination Raman microscopes [73–75] which supply an illumination laser line to scan a sample in question resulting in significantly shorter acquisition time (typically several hundreds times) compared to raster scanning based on point illumination. There, the illumination line, typically generated through a sequence of cylindrical lens, creates some non-homogeneous illumination source, which can affect the spatial distribution of photon counts in a Raman image and negatively impacts the results of the subsequent chemometrics analysis. Despite attempts to remove the consequence of a non-uniform illumination source in Raman data, using scaling techniques such as area normalization under a spectral curve, there exists room to improve further, as demonstrated in the paper. Additionally, non-uniform illumination in Raman microscopy can be caused by various factors such as laser misalignment, poor lens quality, dust, or vignetting effects, and has a negative effect on all types of Raman microscopes. Therefore, techniques that address non-uniform illumination correction are in high demand for the restoration of Raman images.

In this paper is presented an analytical methodology that effectively eliminates non-

uniform illumination in Raman images using the Karhunen-Loeve basis[76]. The method's performance is evaluated using follicular thyroid cancer cells (FTC-133) [10, 12] and normal thyroid cells (Nthy-ori 3-1) as samples in a Raman measurement analysis. Accurate wavenumber calibration is emphasized to avoid potential inaccuracies, including reference sample Raman peak position shifts caused by uneven illumination, and is performed pixel-by-pixel along the line axis. The standard preprocessing protocol [77, 78] recommended in the literature for Raman tabular data is found to be insufficient in correcting intensity variation in Raman data due to non-uniform illumination, as indicated by the existence of a correlation between the spatial coordinates (illumination axis, scanning axis) and the distribution of Raman intensities at different wavenumbers. Therefore, potential misclassification of cells based on their spatial location rather than their actual chemical composition are unavoidable.

We propose a solution to mitigate the issue of intensity variations coming from uneven illumination laser source in Raman images using a random forest regression model[52] in the Karhunen-Loeve basis. Following the position-dependent wavenumber calibration scheme along illumination line (axis), the process involves estimating low-frequency dependencies between the illumination axis and chemical features expressed in the basis, and subtracting these estimations from each chemical feature to minimize unwanted intensity variations along the axis. The same procedure is repeated for the vertical axis (scanning direction) to the illumination line to further minimize intensity variations throughout the images. This process, similar to a detrending technique, assumes that each individual chemical feature in the basis should follow a symmetric distribution. The proposed method is applied after standard preprocessing, and its performance is evaluated through a comparison of chemical homogeneity among single cells from the same phenotype. The results show that this method significantly improves chemical homogeneity between single cells of the same phenotype, and enhances chemical separability between two different phenotypes, FTC-133 and Nthy-ori 3-1, by reducing the risk of misclassification caused by undesired intensity variations.

4.1 Cell culture

In this research, two cell lines were used: FTC-133 (human thyroid follicular carcinoma) as a cancer cell line and Nthy-ori 3-1 (human thyroid follicular epithelial) as a non-cancer cell line. The cells were seeded in a 2mL medium containing DMEM/Ham's F-12 (FUJIFILM Wako Pure Chemical Corporation, 042-30795) for FTC and RPMI1640 (nacalai tesque, 05176-25) for Nthy, along with 10% fetal bovine serum (GE Healthcare, SH30910.03) and 1% penicillin-streptomycin-glutamine (FUJIFILM Wako Pure Chemical Corporation, 161-23201) at a cell number of 2×10^5, on a calcium fluoride substrate (CRYSTRAN LTD, Raman grade CaF_2 CAFP13-0.2). After seeding, the cells were incubated in a CO_2 incubator (5% CO_2 , $37°C$) for 40-48 hours. Before the Raman measurement, the cellular culture medium was replaced with warmed-up Tyrode's buffer solution (145 mM NaCl, 1 mM $CaCl_2$, 1 mM $MgCl_2$, 5.4 mM KCl, 10 mM glucose and 10 mM HEPES with deionized distilled water at a final pH of 7.4) after being rinsed twice with it.

4.2 Line illumination Raman microscope

The Raman images were acquired using a home-built line-illumination Raman microscope[18] equipped with a continuous wave laser at 532 nm (Verdi V18; COHERENT). The power density was set to 3.3 mW/ μm^2 at the sample and the laser was expanded into a line shape using a cylindrical lens, then focused onto the sample through a $\times$40 water immersion objective lens (NA 1.25, CFI Apo 40$\times$WI λ S; Nikon). The Raman photons were backscattered through the same objective lens and collected by a spectrophotometer (MK-300; Bunkoukeiki) after passing through a long-pass edge filter (LP03-532RE-25; Semrock) that eliminates excitation line emission and Rayleigh photons. The Raman photons were dispersed by a 600 L/mm grating(500 nm blaze) and the dispersed transmission was captured by a cooled (-70°C) CCD camera (PIXIS 400 BeXcelon;, Teledyne Princeton Instruments). For each passage the line allows the collection of 400 Raman spectra simultaneously with an exposure time of 5s. To form Raman images, a galvano-mirror is used to scan the sample with the line from left to right. A Raman image was 400×240

pixels totaling 96,000 spectra, each with a spectral range of 182 cm^{-1} to 3,086 cm^{-1} and a size of 910 pixels. In the following, we denote Raman image by a data cube $u(m,n,\nu)$, in which $(m,n,\nu) = (400,240,910)$ in this work. The spectrometer calibration was carried out using ethanol by a spectrometer software.

4.3 Data set characteristics and post-processing

Ten Raman images $\{\hat{u}_i\}$ $(i = 1,..,10)$ (5 FTC-133 and 5 Nthy-ori 3-1) were considered for the analysis. From these 10 images, 60 single cells (28 cells of FTC-133 and 32 cells of Nthy-ori 3-1) were extracted based on manual image segmentation. We preprocessed each individual spectrum belonging to the cell region where the preprocessing runs over all pixels belonging to the defined cell regions. As overall, the sample size of the preprocessed data is 362,593, with 5 labeled FTC-133 and 5 labeled Nthy-ori 3-1.We performed k-means clustering[79, 80] based on the individual spectrum of each Raman image to identify the uniforminty of the proportion of the clusters within the individual single cells. For low dimensional projection of all single cells average spectra, we applied multidimensional scaling (MDS) [46, 47] and Uniform Manifold Approximation and Projection (UMAP) [81, 82].

4.4 Results and discussion

In this section, we first highlight the significant impact of wavenumber calibration along the line axis to detect subtle differences in Raman spectra between human thyroid carcinoma FTC-133 and Nthy-ori-3-1 cell lines. Our explanatory analysis reveals that the use of a standard wavenumber calibration procedure determined with a reference sample measurement independently of positions along illumination line can lead to the emergence of artificial Raman intensity spatial biases. This issue has not been widely acknowledged in the literature. Second, to reduce intensity variations related to uneven illumination in Raman images we propose a preprocessing workflow that introduces some spatial correction in the principal component basis. This approach effectively reduces non-uniform inten-

sity profiles in Raman images and enhances the accurate differentiation of Raman spectra between FTC-133 and Nthy-ori 3-1 cell lines. Here, I will discuss the impact of three preprocessing schemes in details that was explained in the Chapter 3.

4.4.1 Applications of the position-dependent wavenumber calibration and the detrending scheme

Fig. 4.1 presents a series of visualization and descriptive statistics estimated on representative Raman images with and without position-dependent wavenumber calibration and with the detrending scheme on top of the calibration. Panel (A) shows an uncalibrated Raman image, while panel (B) shows the same image with position-dependent wavenumber calibration, which significantly reduces the artificial spatial correlation of the image. Panel (C) shows the Raman image with the detrending scheme on top of the calibration. Panels (D) and (E) show Pearson correlation coefficients (r) between the Raman image at each individual wavenumber and the illumination axis ζ and the scanning axis ξ, respectively. Panel (F) shows the averaged with one standard deviation Raman spectrum for the cell region with the three preprocessings. Fig. 4.1D, without wavenumber calibration, shows high positive correlation at certain Raman shifts, high negative correlation at other Raman shifts, and weak correlation for some Raman shifts. The sign of correlation coefficient is dependent on the definition of the coordinate system. Suppose that a Pearson correlation coefficient is positive about +0.8. This is equally possible to be -0.8 if one inverts the axis from positive to negative in the definition of the coordinate system. However, note that the relative relationship, such that some Raman intensities correlate along one direction (e.g., positively) but the others do along the inverse direction (e.g., negatively) to the chosen axis, holds once the coordinate system is fixed. This spatial correlation is significantly reduced by the position-dependent calibration strategy indicating an apparent wavenumber drift along the illumination axis, which could be attributed to chromatic aberration or changes in physical properties resulting from laser light power variation along this axis. Along the scanning axis some correlation pattern also exist for data without wavenumber calibration as observed Fig. 4.1E but with lower amplitude than observed for the illumination axis.

Similarly, same correlation analysis for all 5 FTC and 5 Nthy Raman images are shown in Fig. 4.2-Fig. 4.5.

4.4.2 Classifications of FTC-133 and Nthy-ori 3-1 based on Raman images

To evaluate the quality of the three different preprocessing schemes including with/without the position-dependent wavenumber calibration and the detrending scheme on the top of the position-dependent wavenumber calibration, on the Raman images, a comparative analysis of the classification performance of the two cell lines was conducted. Additionally, to provide a visual representation of the effect of the three preprocessing strategies on the data, the average single cell Raman spectra are projected into a low-dimensional space. Fig. 4.7A, B, and C visualizes the projection of sixty average single cell spectra in a low-dimensional space by performing multidimensional scaling (MDS) [46, 47] based on the distance matrix (Fig. 4.6). This low dimensional representation manifests that the detrending scheme clearly enhances the differentiability between FTC-133 and Nthy-ori 3-1, as shown in Fig. 4.7C (c.f., Fig. 4.7A and B) and in linear PCA projection shown in Fig. 4.11 as well.(See also Fig. 4.12) for the nonlinear projection, Uniform Manifold Approximation and Projection (UMAP) [81, 82]). It is revealed that the enhanced differentiability by the detrending scheme on the top of the position-dependent wavenumber calibration is statistically ensured, free from the choice of 2D linear basis of MDS or by using 2D nonlinear embedding algorithm like UMAP. Fig. 4.7D and E shows the box-and-whisker plot of 25 cross-validated accuracies of random forest classifier (RFC) [83, 84] models in predicting of FTC-133/Nthy-ori 3-1 for the three different preprocessing schemes. That is, for each preprocessing, a pair of two images of FTC-133 and Nthy-ori 3-1 were randomly chosen 25 times as test images to estimate the classification accuracy while the remaining 4 FTC-133 and 4 Nthy-ori 3-1 images were used to train RFC. The RFC creates an ensemble of 100 decision trees on different subset of the training data on Raman spectra coming from the training set, and predict class (FTC-133 or Nthy-ori 3-1) membership of unseen Raman spectra from the test set based on the majority class voting of the 100 decision tress.

Fig. 4.7D shows the RFC accuracy when considering average single-cell Raman spectra, while Fig. 4.7E is obtained by considering all the spectra belonging to cells. From these figures, it is evident that a proper wavenumber calibration adapted for line-illumination microscopes and/or a detrending scheme is essential to stabilize the performance of the classifiers. Indeed, the average RFC accuracy increases progressively from uncalibrated data to detrended data, while the standard deviation of the accuracy decreases. This trend emphasizes that our preprocessing method improves the stability of RFC classifier by reducing the number of outliers. Similar trend exists for AUC and f1 score calculations are shown in Fig. 4.8 and Fig. 4.9 respectively. Moreover, Fig. 4.10 shows the box-and-whisker plot of 25 cross-validated accuracies by CNN for the three different preprocessing schemes based on pixelwise spectra. The detrended scheme exhibits better performance compared to both the uncalibrated and calibrated schemes.

4.4.3 Visualization of spectral stability via a cluster analysis

Fig. 4.13A, B, and C depict the results of *k*-means clustering maps with 5 clusters for a representative Raman image of human follicular thyroid carcinoma cell line FTC-133 for the three different preprocessing schemes. Here, the *k*-means clusterings were performed independently for each preprocessing strategy. To compare visually between the three sets of clusters, the cluster indices for each scheme were reordered in each image so that the Euclidean distance between the centroid (corresponding to the median in spectral space) of each cluster computed for three different preprocessed Raman images is minimized by rearranging the index of cluster for each image (See the distance matrix between the rearranged clusters of the three data sets in Fig. 4.14). In Fig. 4.14, the corresponding clusters between the three preprocessed images were able to be identified rather straightforwards. The corresponding cluster population ratios can be seen in Figures 4.13D-F. One can see in Figures 4.13D-F that the population of the clusters within individual cells tend to be relatively more diverse without using the detrending scheme. For example, the proportion of cluster 3 (megenta) is relatively very high for cell 3 compared to other cells, which would suggest the manifestation of a phenotypic difference. In turn, the proportion of cluster 3

for cell 3 as well as other clusters are relatively less diverse across different single cells for the detrending scheme compared to the other two preprocessing. Similarly, the results of the same analyses for all ten Raman images are given in Fig. 4.15-Fig. 4.18. We then emphasize that the detrending scheme reduce the variation of cluster proportion between single cells of a Raman image. This is demonstrated Fig. 4.13G by showing a comparative single-cell pairwise spectral cluster proportion Euclidean distance distribution for the three preprocessing strategies via a box-and-whisker plot. Figure 4.13G manifests that the detrending scheme naturally provides statistically consistent population distributions for each single cell within the same image.

Fig. 4.19 summarizes the average Raman spectra variation of individual cells obtained from 10 Raman images in terms of the three preprocessing strategies. Panels (A) and (B) show the average spectra of 28 Nthy-ori 3-1 and 32 FTC-133 single cells, respectively, for uncalibrated data. Panels (C) and (D) present the corresponding average spectra of the same number of cells after incorporating the position-dependent wavenumber calibration. Lastly, Panels (E) and (F) display the average spectra of the mentioned single cells after both the position-dependent wavenumber calibration and the detrending scheme have been implemented. The box-and-whisker plot for variation of Raman intensities in Fig. 4.19G and Fig. 4.19H for Nthy-ori 3-1 and FTC-133, respectively, shows reductions in the variance of the average Raman spectra as the three preprocessing strategies. We interpret that this observed reduction of variance through the utilization of the 3 preprocessing strategies is related to a minimization of unwanted variations coming from instrumental or experimental factors. This highlights the importance of developing proper preprocessing strategies to obtain results in Raman imaging experiments consistent with enhanced differentiability of the phenotypic differences. Sixty single cells average spectra of Nthy-ori 3-1 and FTC-133 together is shown in Fig. 4.20. At the pixel level, we observed very clearly that intensity distribution is compact by my detrending scheme compare to standard preprocessing schemes in Fig. 4.21.

4.4.4 Confusion matrix and predicted probability

In Fig. 4.22, we found that 3 FTC spectra were misclassified and these 3 spectra located far from the FTC group in UMAP space. Moreover, average spectra of these 3 cells are different than other 3 cells in case of standard preprocessing without wavenumber calibration. However, in Fig. 4.23, we observed that all 6 FTC spectra were correctly classified but 3 spectra located far from the FTC group in UMAP space shows very marginal probability around 0.5 in case of standard preprocessing with position-dependent wavenumber calibration. But in case of the detrending scheme after the implementation of position-dependent wavenumber calibration (Fig. 4.24), 6 cells were closer in UMAP space and predicted probability becomes higher. Moreover,average spectra of these 6 cells are nearly similar. A similar finding was observed for the another test pair (Nthy4 and FTC2) are shown in the Fig. 4.25-Fig. 4.27.

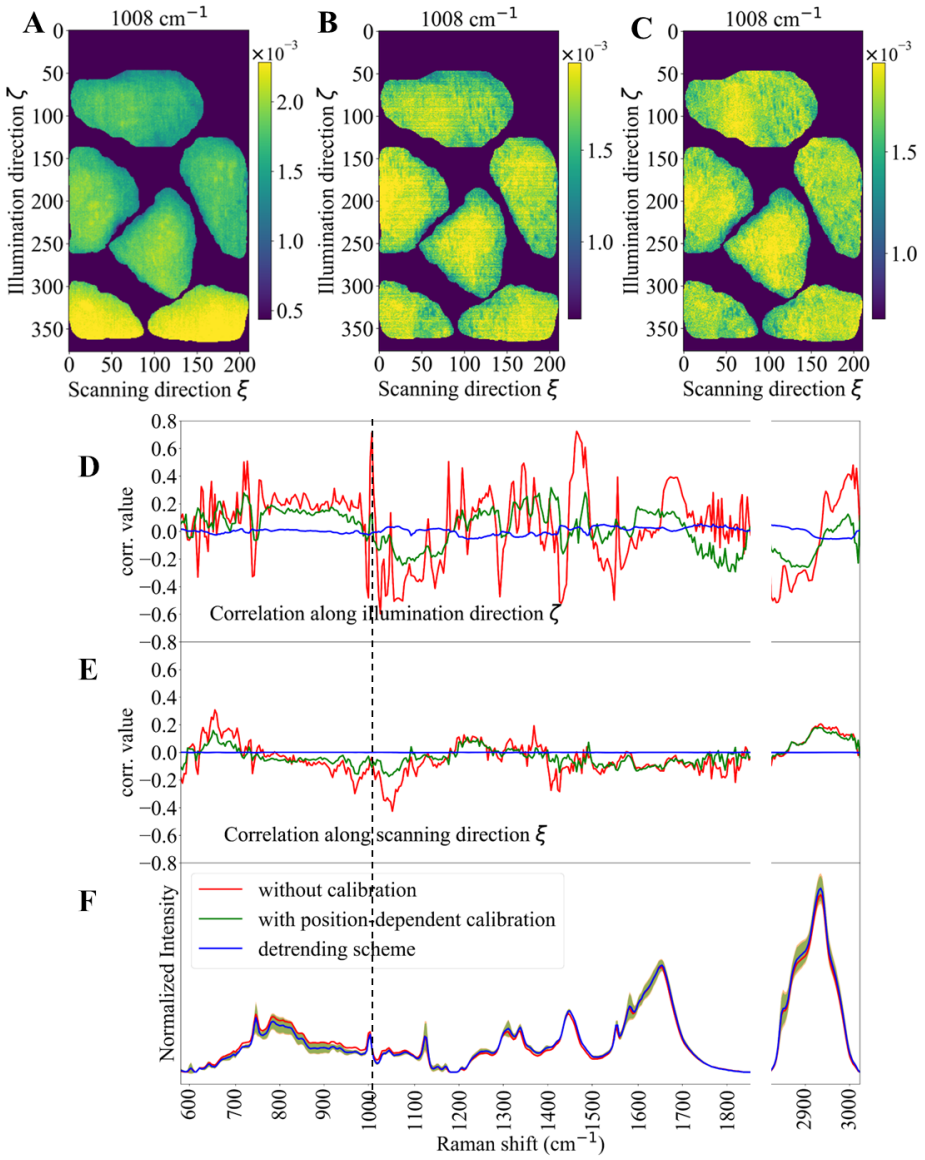


Fig. 4.1. (A)-(C) The Raman intensity distribution at 1008 cm^{-1} (dashed vertical line) in the space domain of FTC-133(#2): (A) after standard preprocessing without wavenumber calibration, (B) after standard preprocessing with position-dependent wavenumber calibration, (C) after the detrending scheme applied on the top of position-dependent wavenumber calibration. (D)-(E) The Pearson correlation coefficients between the Raman images at each wavenumber acquired by the three preprocessings: (D) the illumination axis coordinate, (E) the scanning axis coordinate. (F) The average Raman spectra over all cell regions, with three different preprocessings. Note that the silent region at wavenumbers 1,880-2,805 cm^{-1} is omitted and replaced by a small gap.

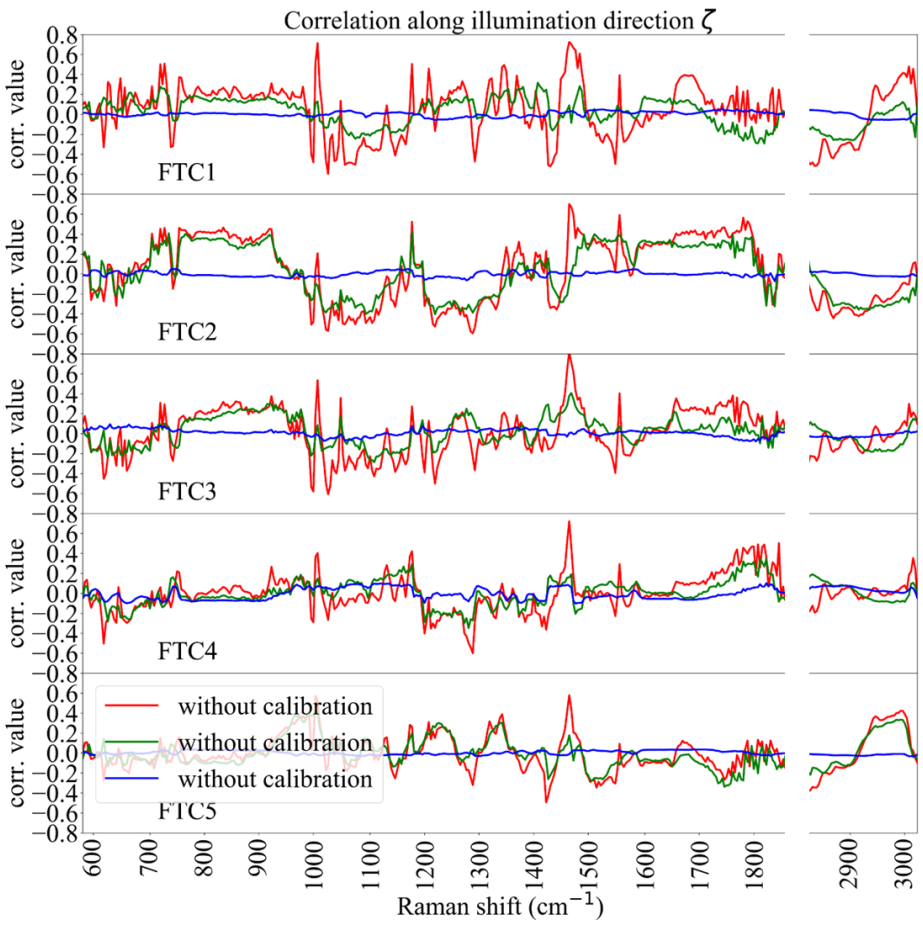


Fig. 4.2. The Pearson correlation coefficients between the Raman images at each wavenumber acquired by the three preprocessings along the illumination axis coordinate for 5 FTC Raman images.

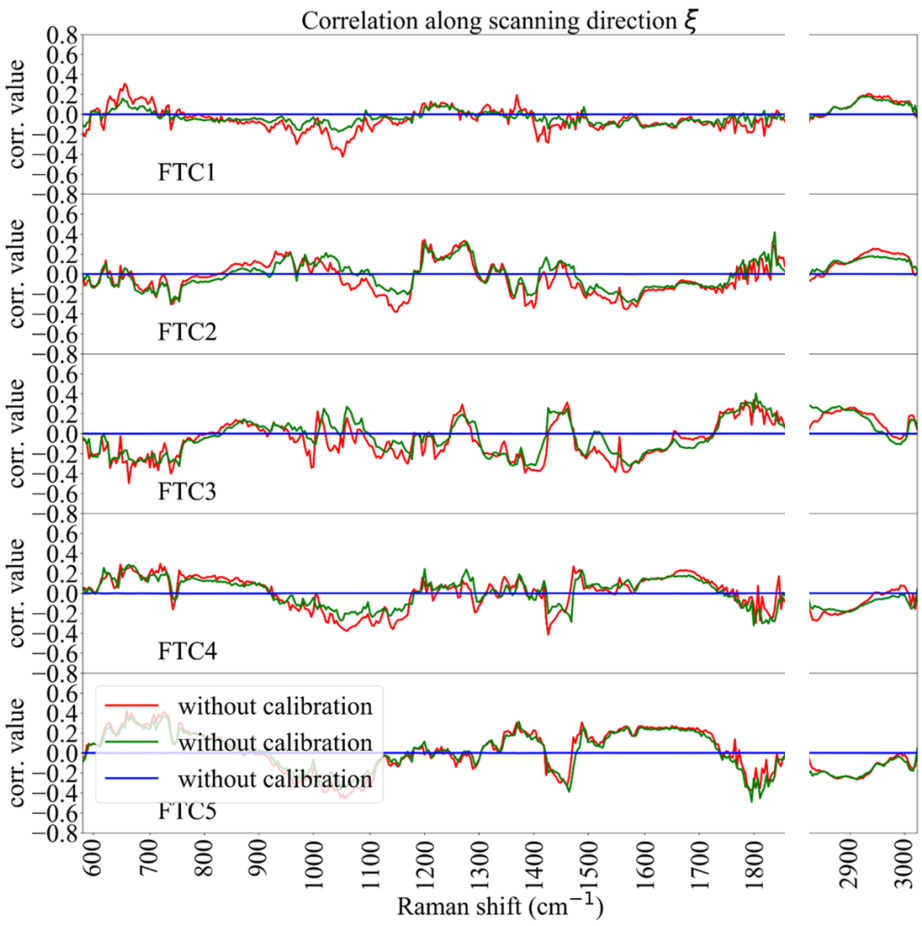


Fig. 4.3. The Pearson correlation coefficients between the Raman images at each wavenumber acquired by the three preprocessings along the scanning axis coordinate for 5 FTC Raman images.

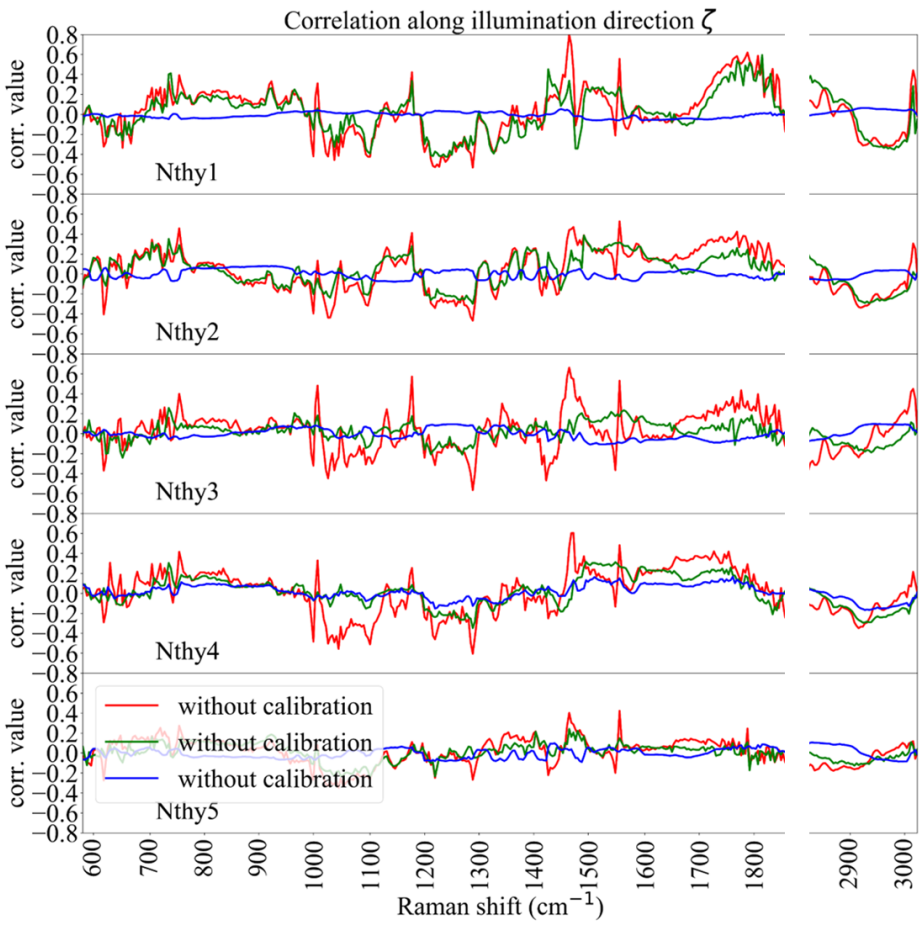


Fig. 4.4. The Pearson correlation coefficients between the Raman images at each wavenumber acquired by the three preprocessings along the illumination axis coordinate for 5 Nthy Raman images.

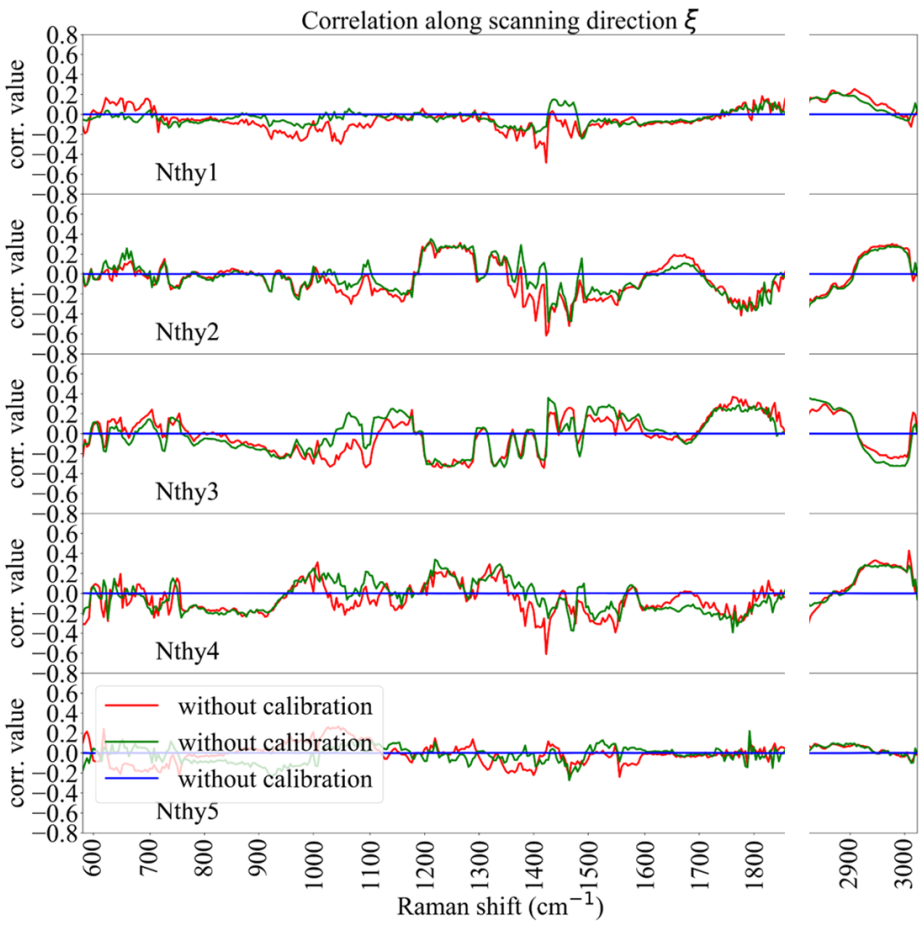


Fig. 4.5. The Pearson correlation coefficients between the Raman images at each wavenumber acquired by the three preprocessings along the scanning axis coordinate for 5 Nthy Raman images.

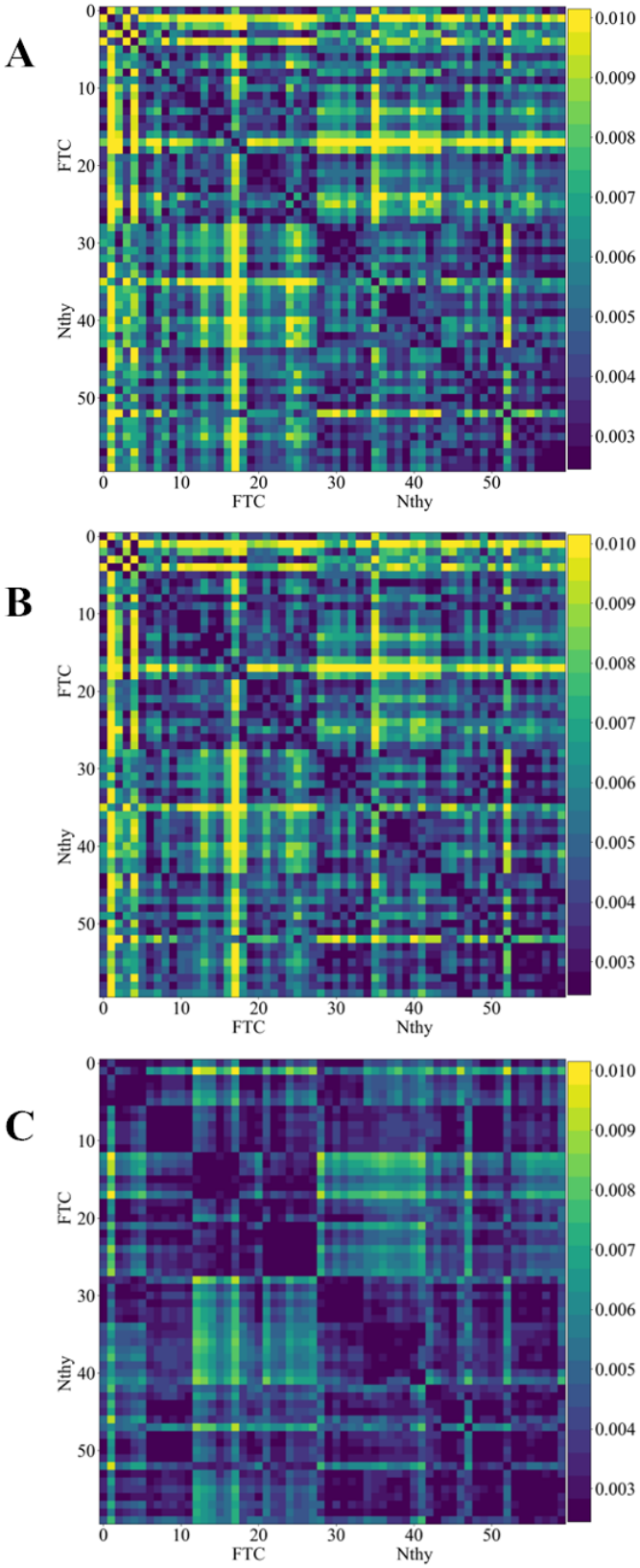


Fig. 4.6. The distance matrix between averaged single cell spectra of sixty cells (28 cells of FTC-133 and 32 cells of Nthy-ori 3-1): (A) without wavenumber calibration. (B) the position-dependent wavenumber calibration. (C) the detrending scheme.

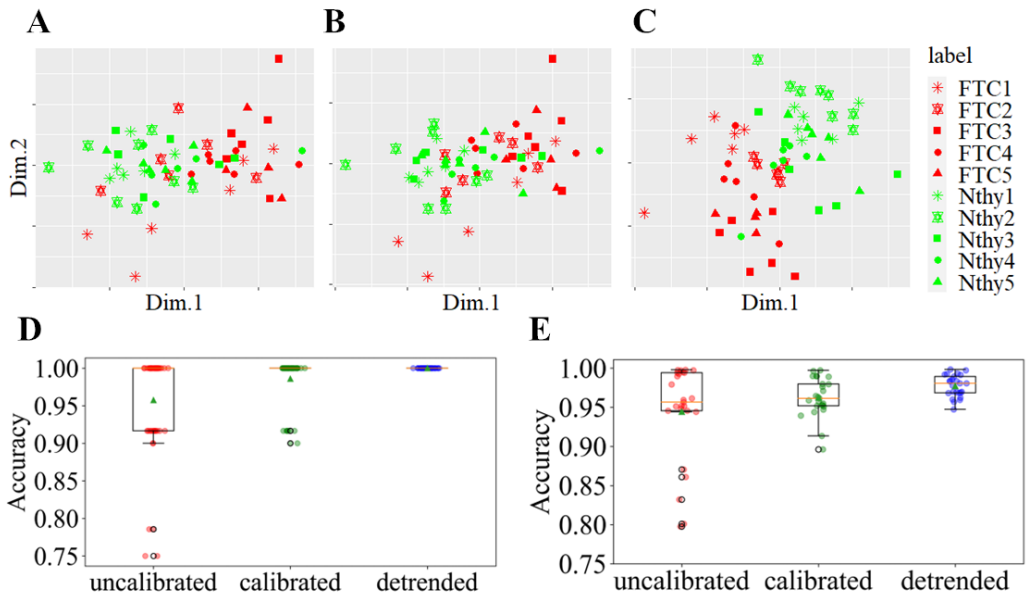


Fig. 4.7. (A)-(C) The multi-dimensional scaling (MDS) projection of the ten Raman images including sixty single cells in total: (A) standard preprocessing without wavenumber calibration (B) the position-dependent wavenumber calibration. (C) the detrending scheme. (D)-(E) The box-and-whisker plot of test accuracy in the prediction of FTC-133/Nthy-ori 3-1 for three different preprocessing schemes of 25-fold cross validation: (D) based on single cell average spectra (E) based on pixelwise spectra.

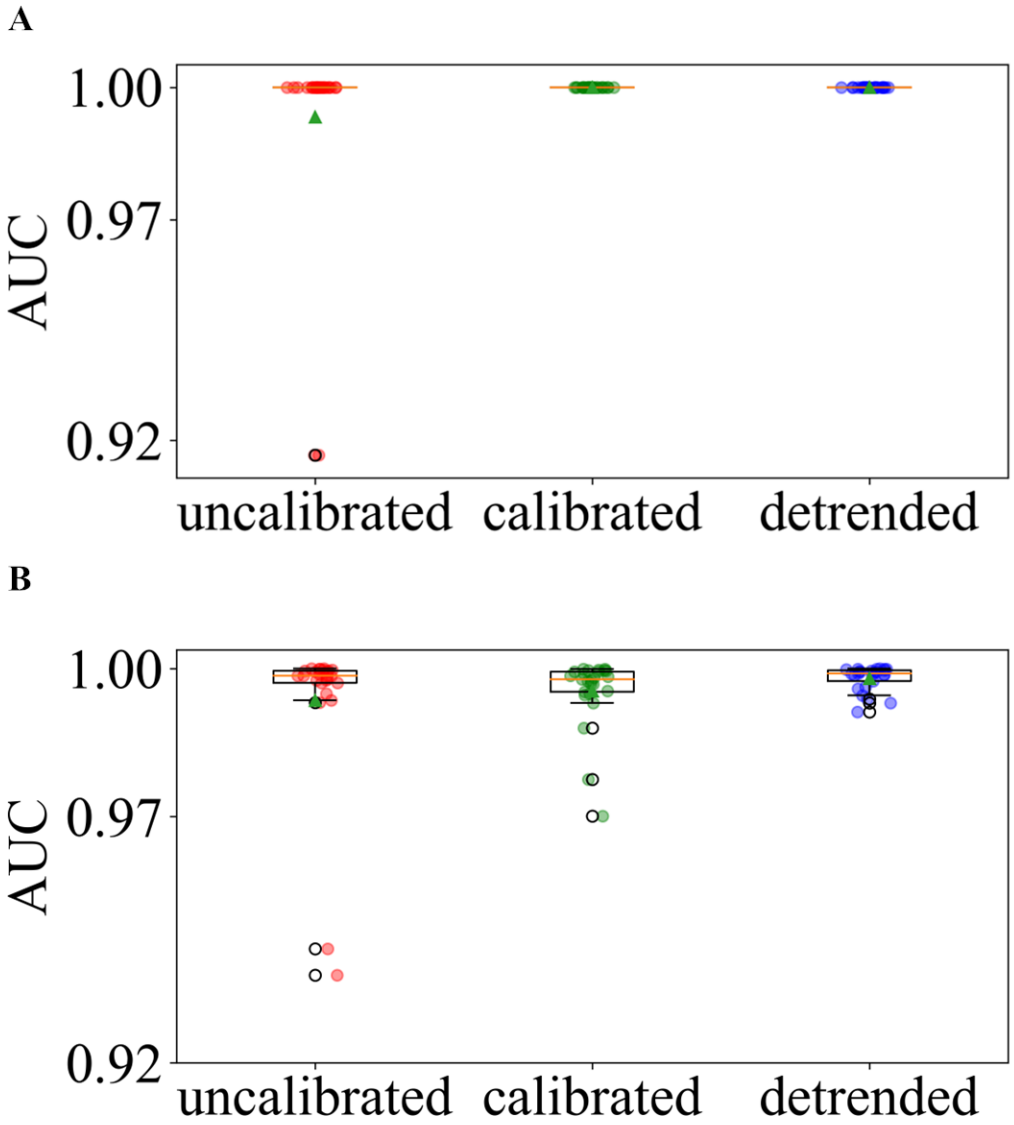


Fig. 4.8. (A)-(B) The box-and-whisker plot of area under curve (AUC) in the prediction of FTC-133/Nthy-ori 3-1 for three different preprocessing schemes of 25-fold cross validation: (A) based on single cell average spectra (B) based on pixelwise spectra.

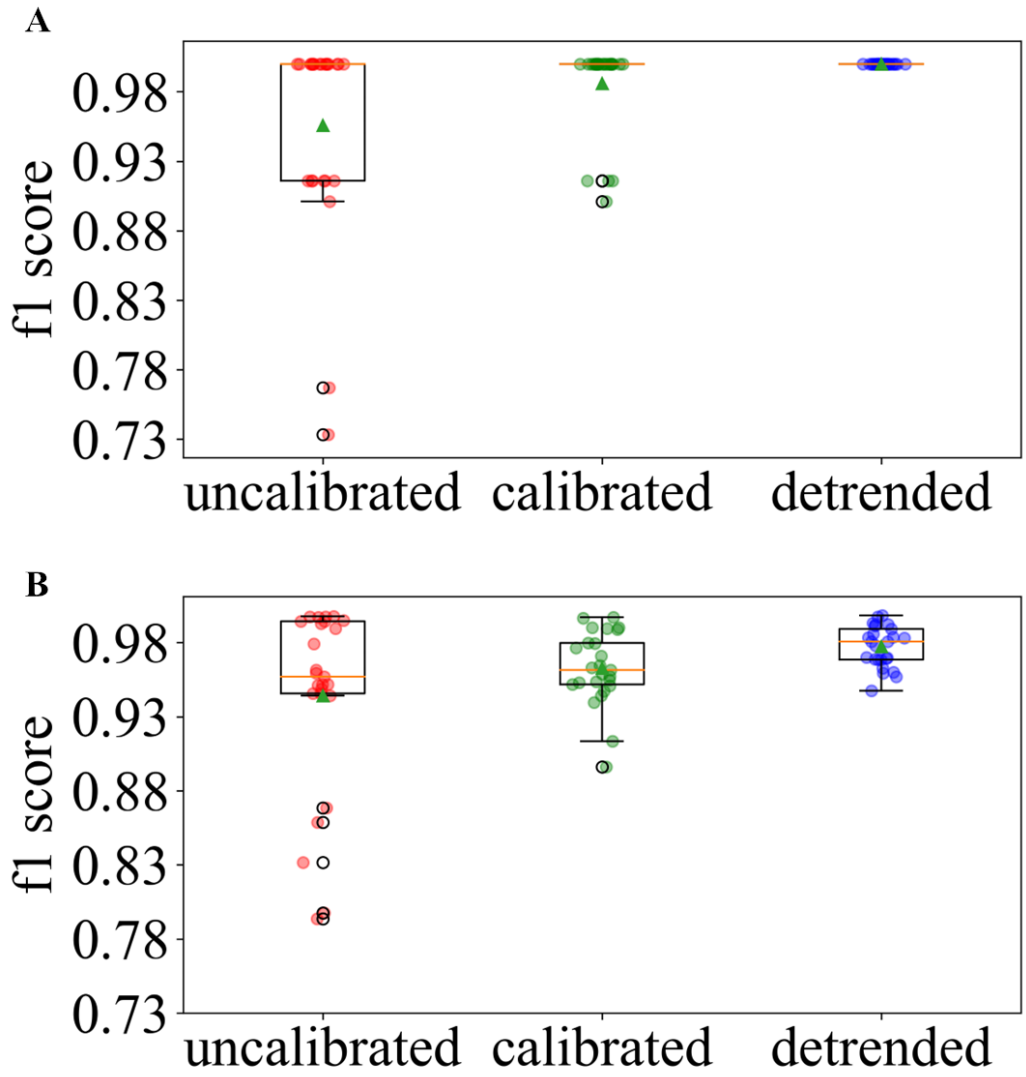


Fig. 4.9. (A)-(B) The box-and-whisker plot of f1 score in the prediction of FTC-133/Nthy-ori 3-1 for three different preprocessing schemes of 25-fold cross validation: (A) based on single cell average spectra (B) based on pixelwise spectra.

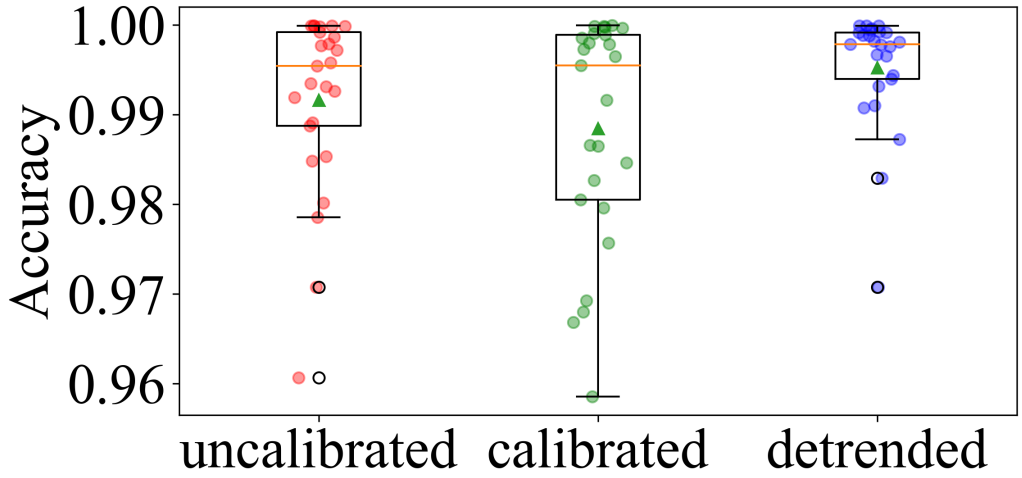


Fig. 4.10. The box-and-whisker plot of accuracy in the prediction of FTC-133/Nthy-ori 3-1 for three different preprocessing schemes of 25-fold cross validation by CNN based on pixelwise spectra.

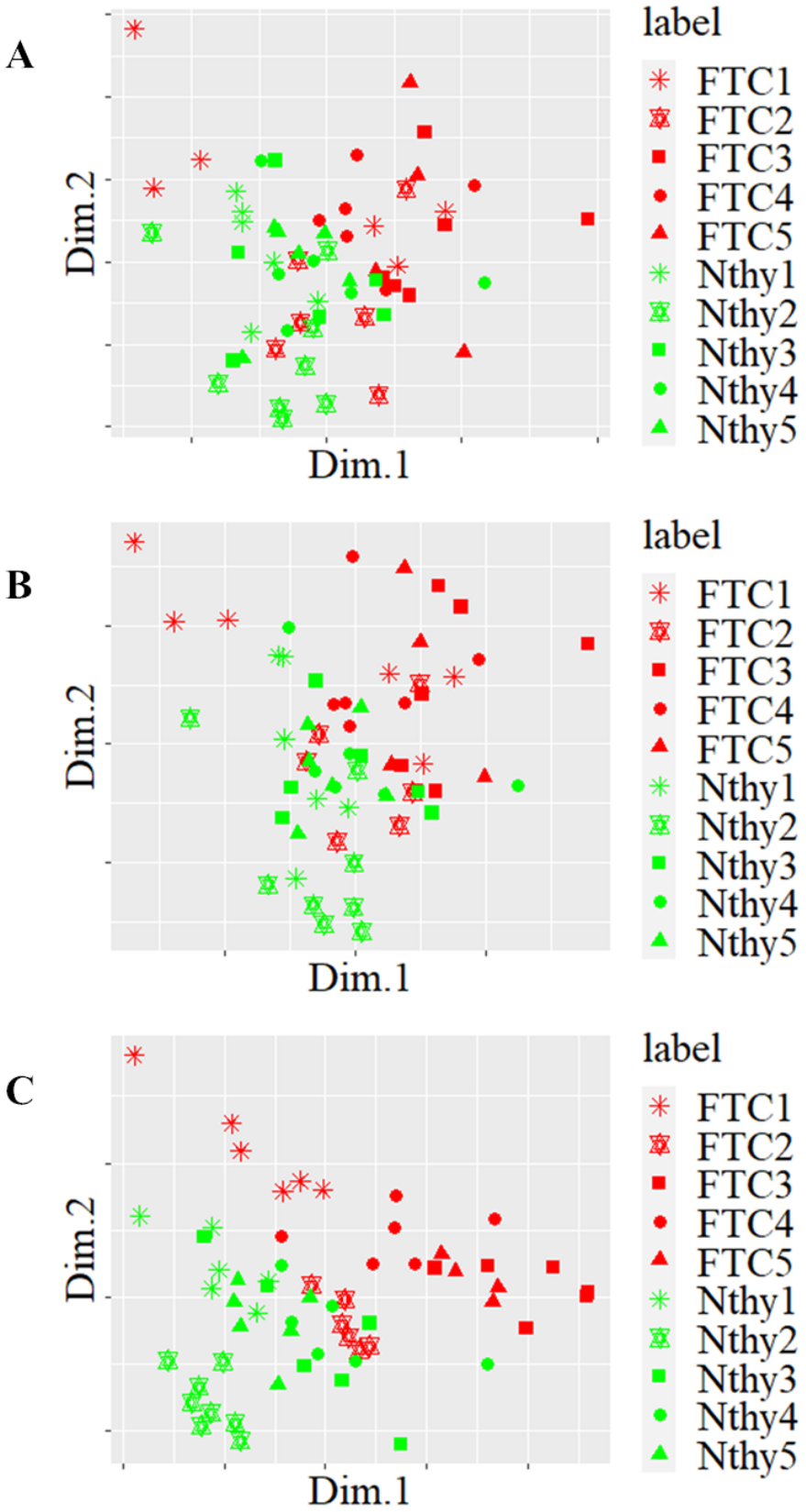


Fig. 4.11. PCA projection of averaged single cell spectra of sixty cells. (A) without wavenumber calibration. (B) the position-dependent wavenumber calibration. (C) the detrending scheme.

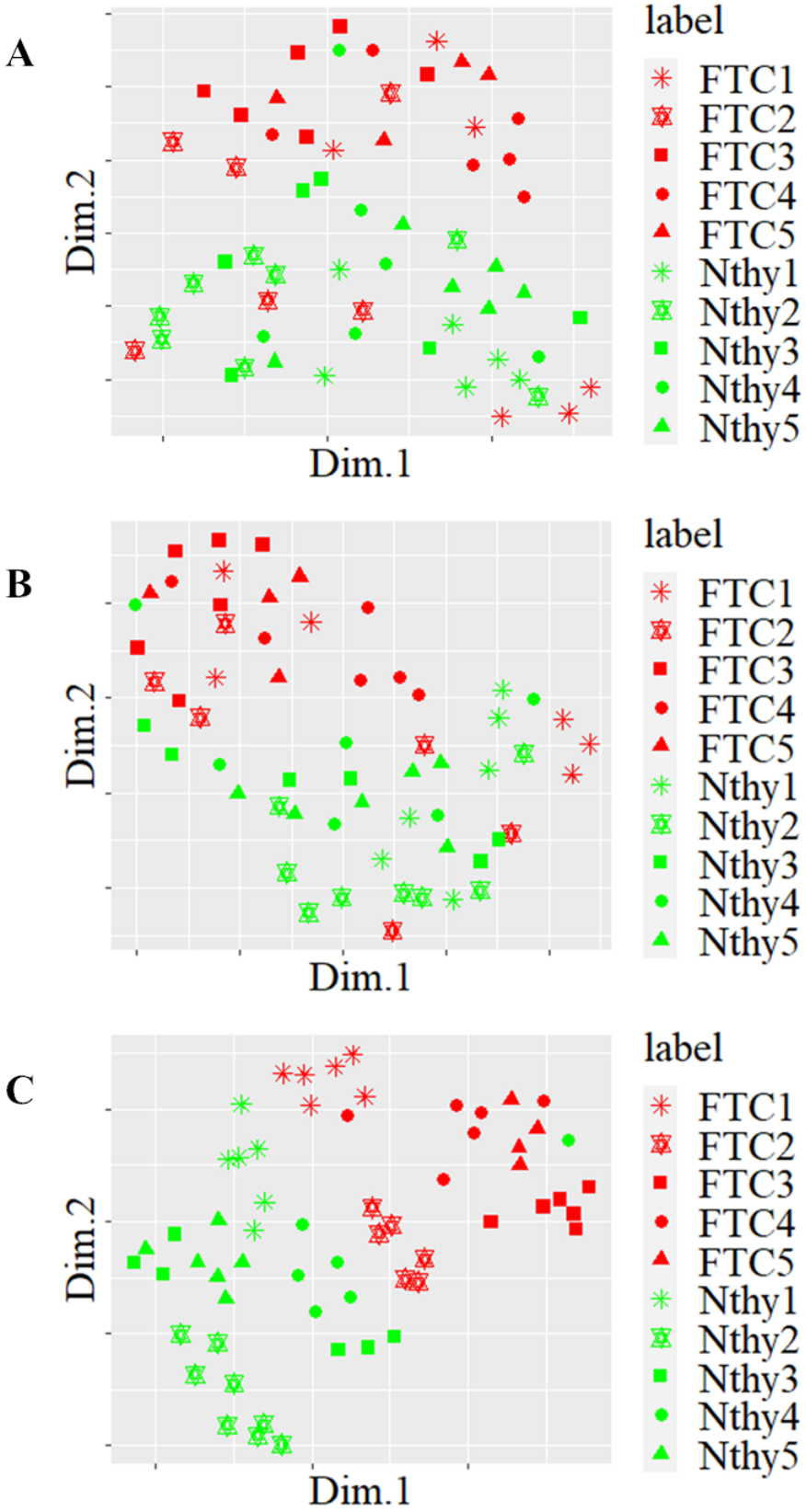


Fig. 4.12. An UMAP projection of averaged single cell spectra of sixty cells. (A) without wavenumber calibration. (B) the position-dependent wavenumber calibration. (C) the detrending scheme.

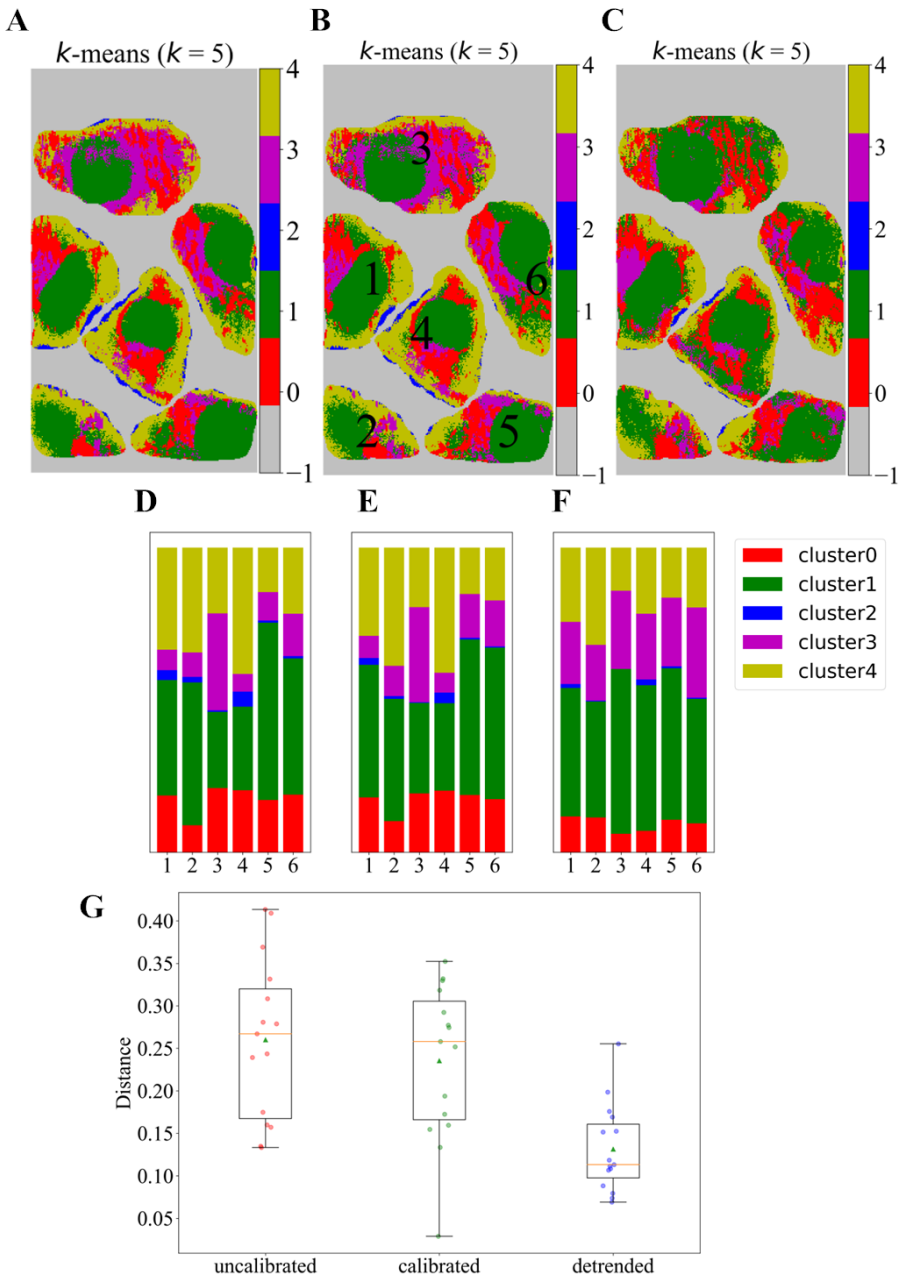


Fig. 4.13. (A)-(C) The k-means clustering maps with $k = 5$ for individual Raman spectra in the Raman image for a representative FTC-133(#2): (A) standard preprocessing without wavenumber calibration (B) the position-dependent wavenumber calibration. (C) the detrending scheme based on random forest regression. (D)-(F) The relative populations of the clusters within each single cell for FTC-133(1): (D) standard preprocessing without wavenumber calibration (E) the position-dependent wavenumber calibration. (F) the detrending scheme. (G) The dependence of the diversity measure of cluster distributions within individual single cells on the kinds of the three preprocessing schemes.

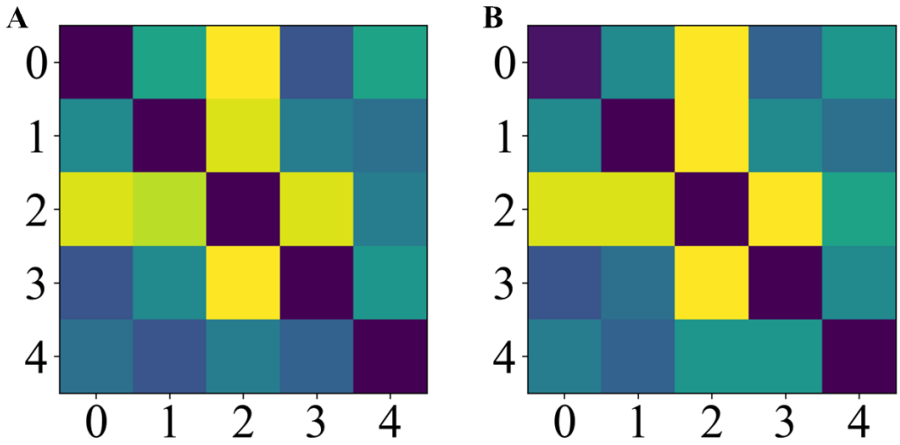


Fig. 4.14. The distance matrix between the centroid of each cluster obtained for the Raman image of FTC-133(1):(A) with the position-dependent wavenumber calibration vs without wavenumber calibration and (B) with the position-dependent wavenumber calibration vs the detrending scheme.

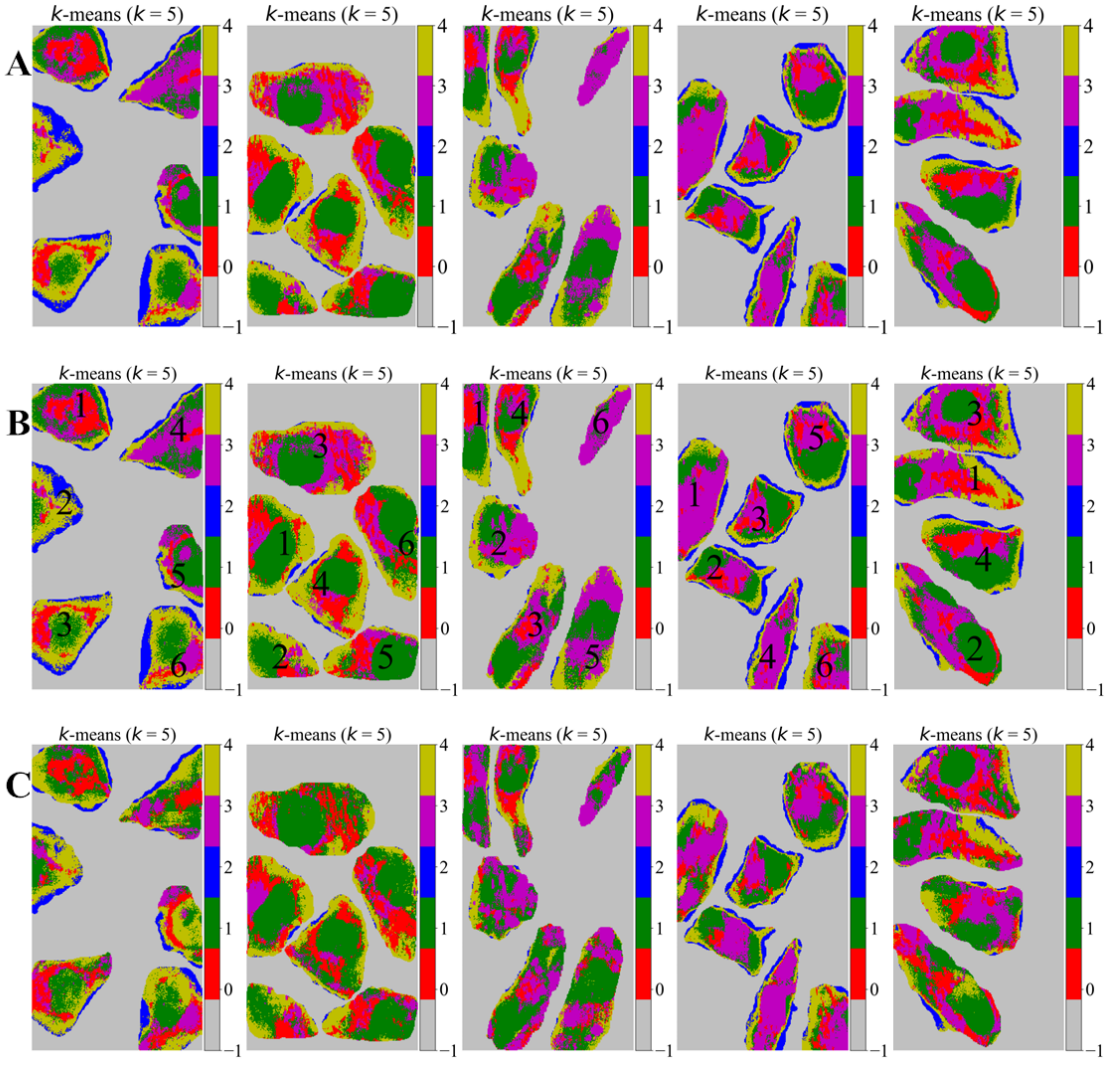


Fig. 4.15. The k-means clustering maps with $k = 5$ for individual Raman spectra of 5 FTC Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

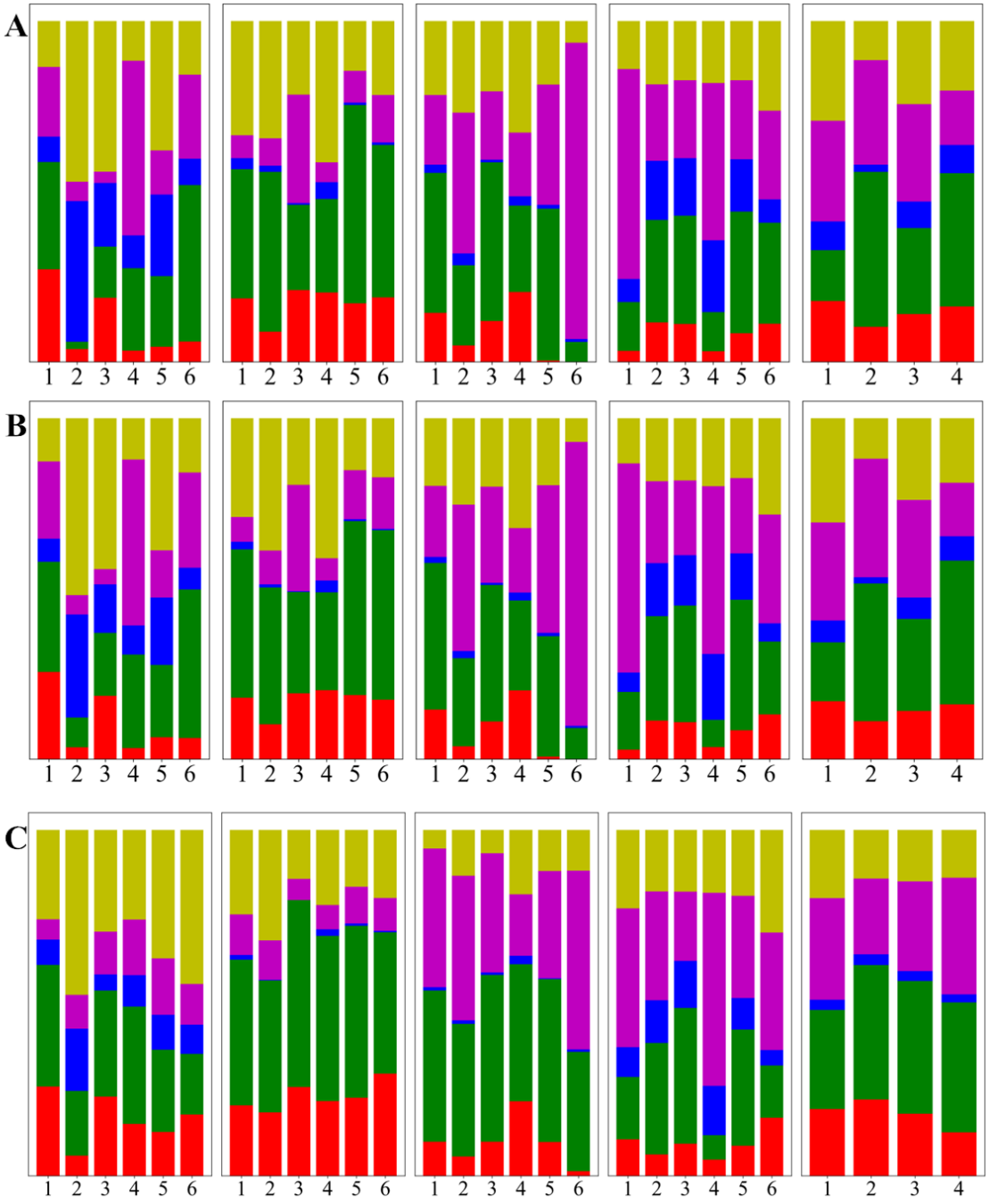


Fig. 4.16. The relative populations of the clusters within each single cell for 5 FTC Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

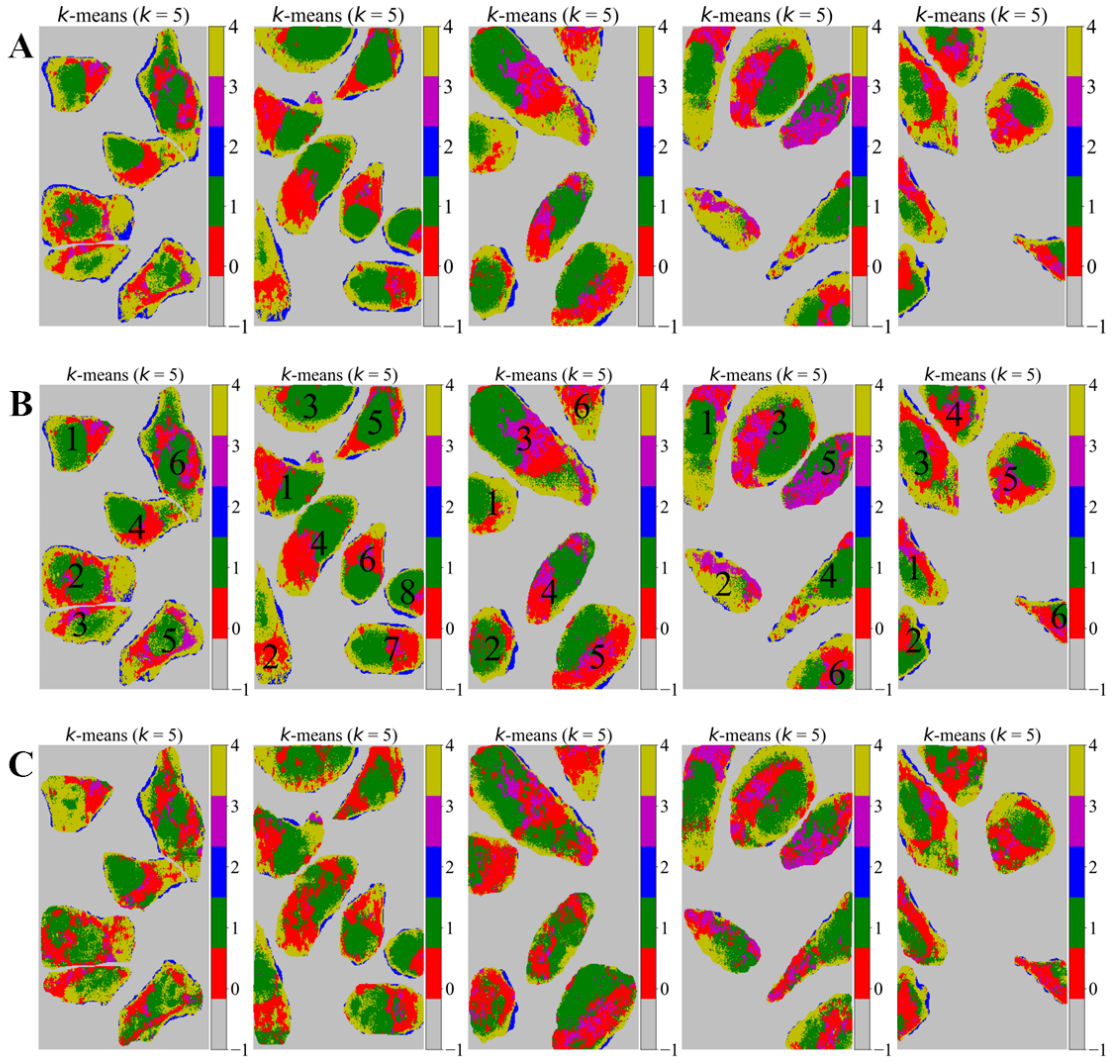


Fig. 4.17. The k-means clustering maps with $k = 5$ for individual Raman spectra of 5 Nthy Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

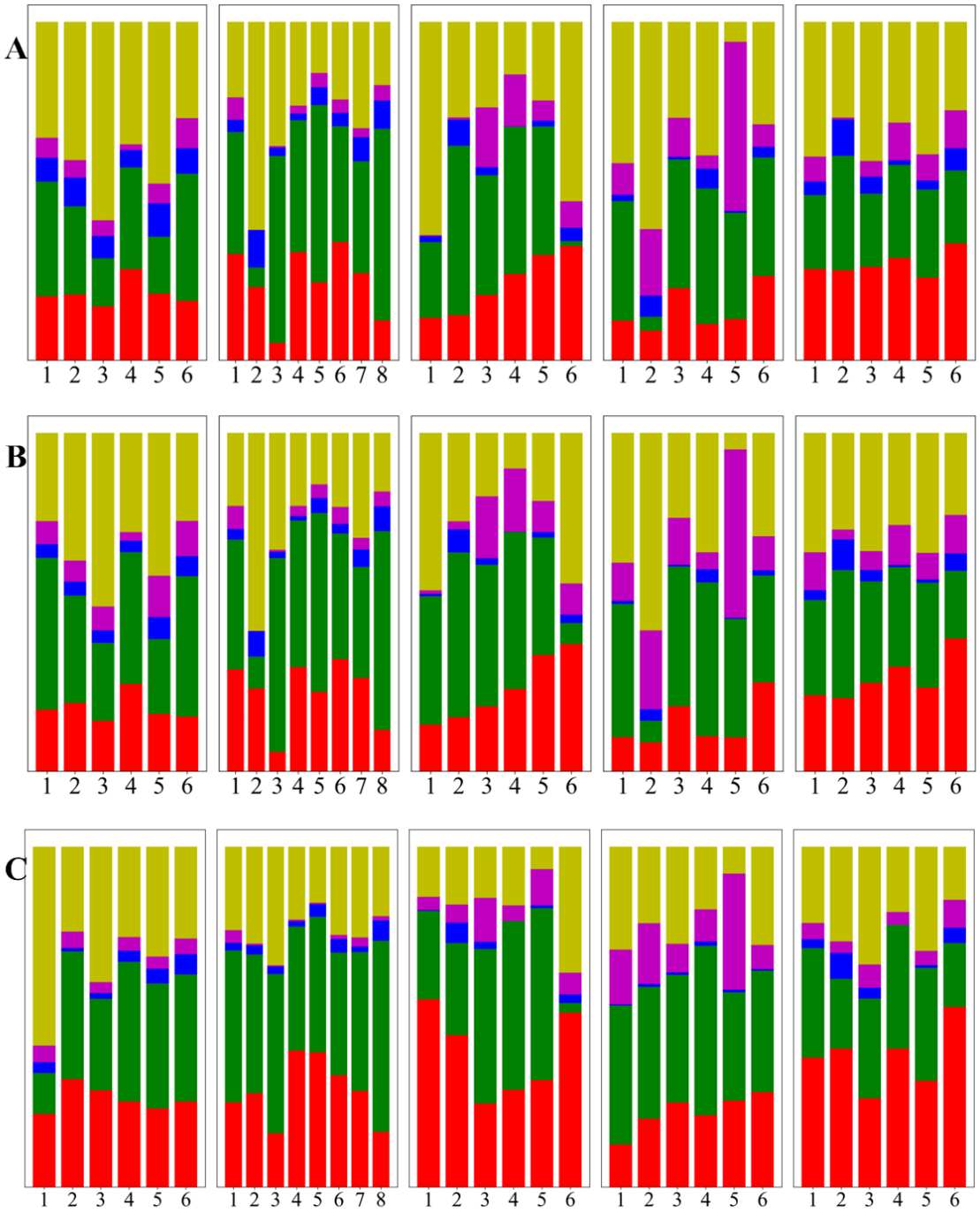


Fig. 4.18. The relative populations of the clusters within each single cell for 5 Nthy Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

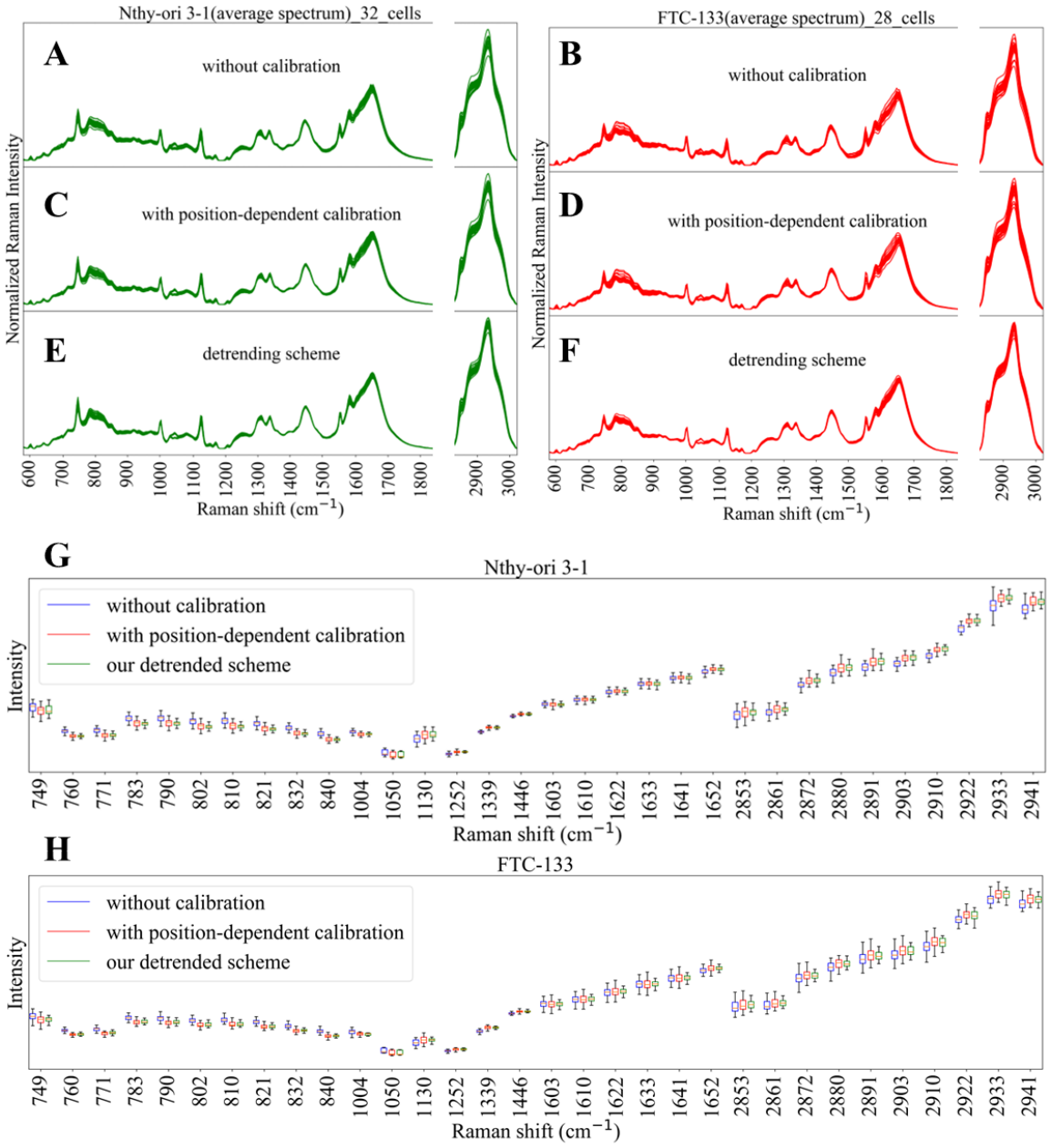


Fig. 4.19. Average spectra of 32 Nthy-ori 3-1 and 28 FTC-133 cells. (A-B) standard preprocessing without wavenumber calibration, (C-D) standard preprocessing with position-dependent wavenumber calibration, (E-F) the detrending scheme after the implementation of position-dependent wavenumber calibration. (G) The box-and-whisker plot for variation of Raman intensities for Nthy-ori 3-1. (H) The box-and-whisker plot for variation of Raman intensities for FTC-133.

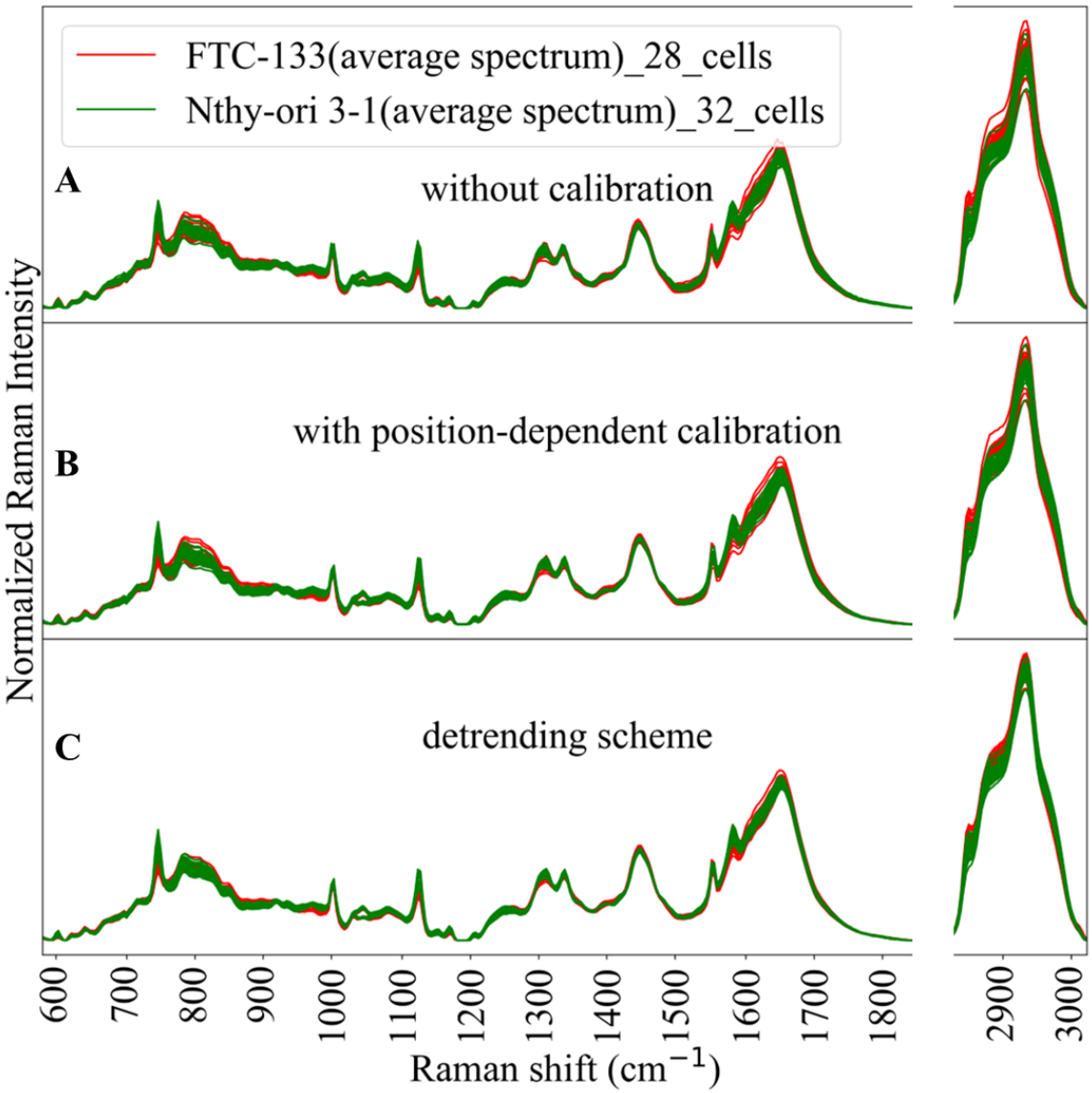


Fig. 4.20. (A-C) Average spectra of 60 cells (32 Nthy-ori 3-1 and 28 FTC-133 cells): (A) standard preprocessing without wavenumber calibration, (B) standard preprocessing with position-dependent wavenumber calibration, (C) the detrending scheme after the implementation of position-dependent wavenumber calibration.

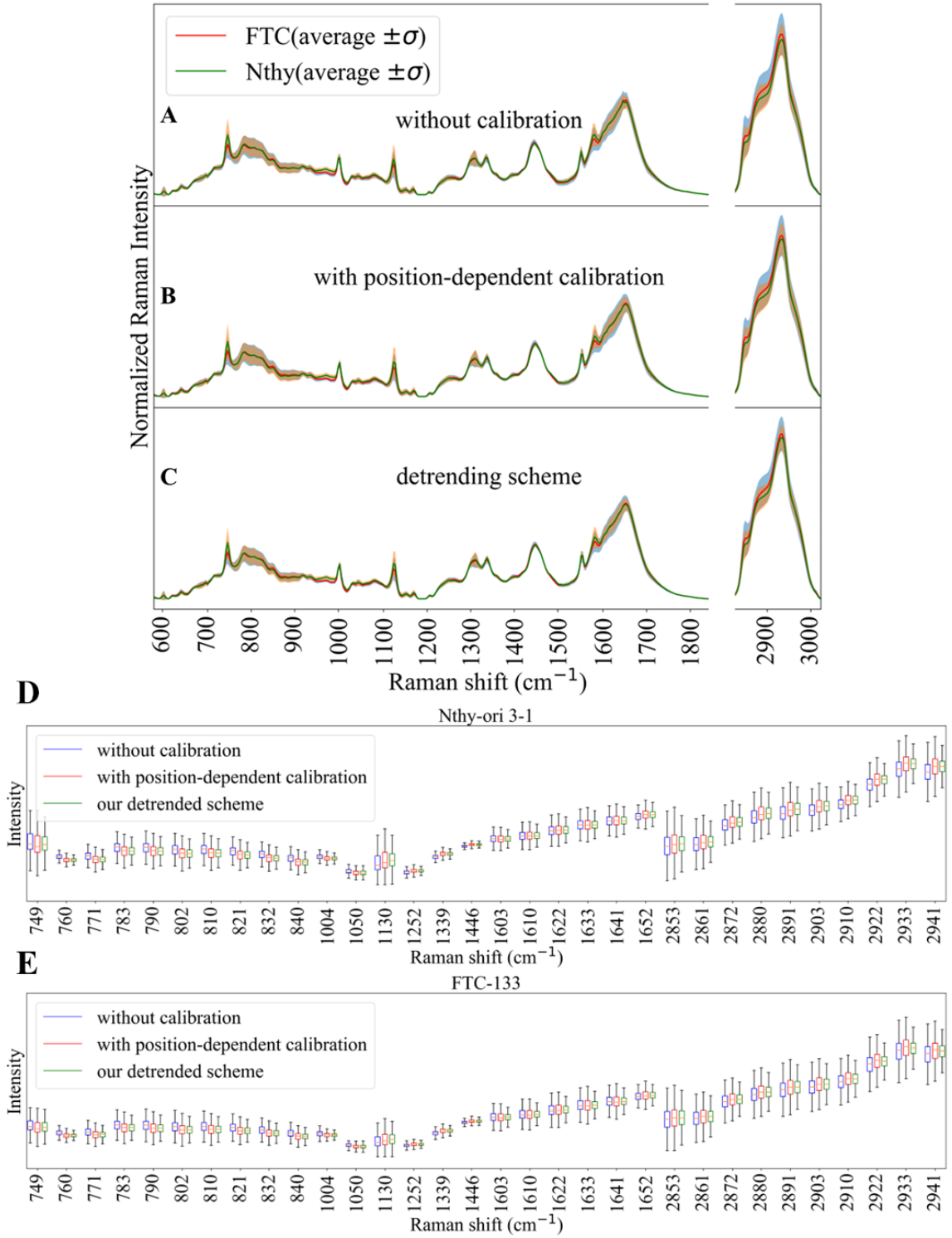


Fig. 4.21. (A-C) Average spectra with one standard deviation of all individual spectra of 32 Nthy-ori 3-1 and 28 FTC-133 cells: (A) standard preprocessing without wavenumber calibration, (B) standard preprocessing with position-dependent wavenumber calibration, (C) the detrending scheme after the implementation of position-dependent wavenumber calibration. (D) The box-and-whisker plot for variation of Raman intensities for Nthy-ori 3-1. (E) The box-and-whisker plot for variation of Raman intensities for FTC-133.

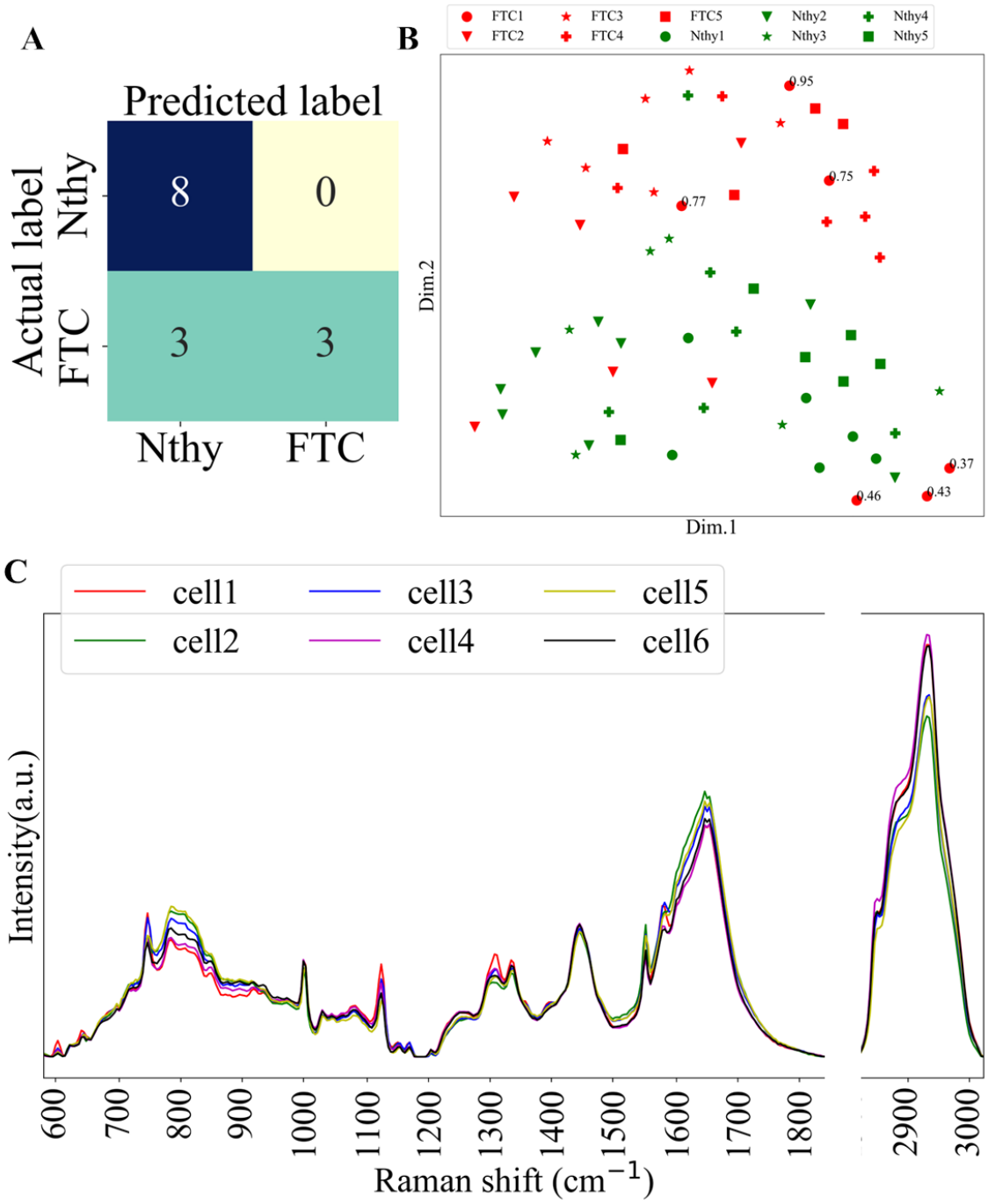


Fig. 4.22. For standard preprocessing without wavenumber calibration: (A) Confusion matrix of test set (Nthy2 and FTC1) (B) predicted probability of 6 cells of FTC1 image in UMAP projection (C) average spectrum of 6 cells of FTC1 image. Note that remaining 8 images (4 Nthy and 4 FTC) were train set.

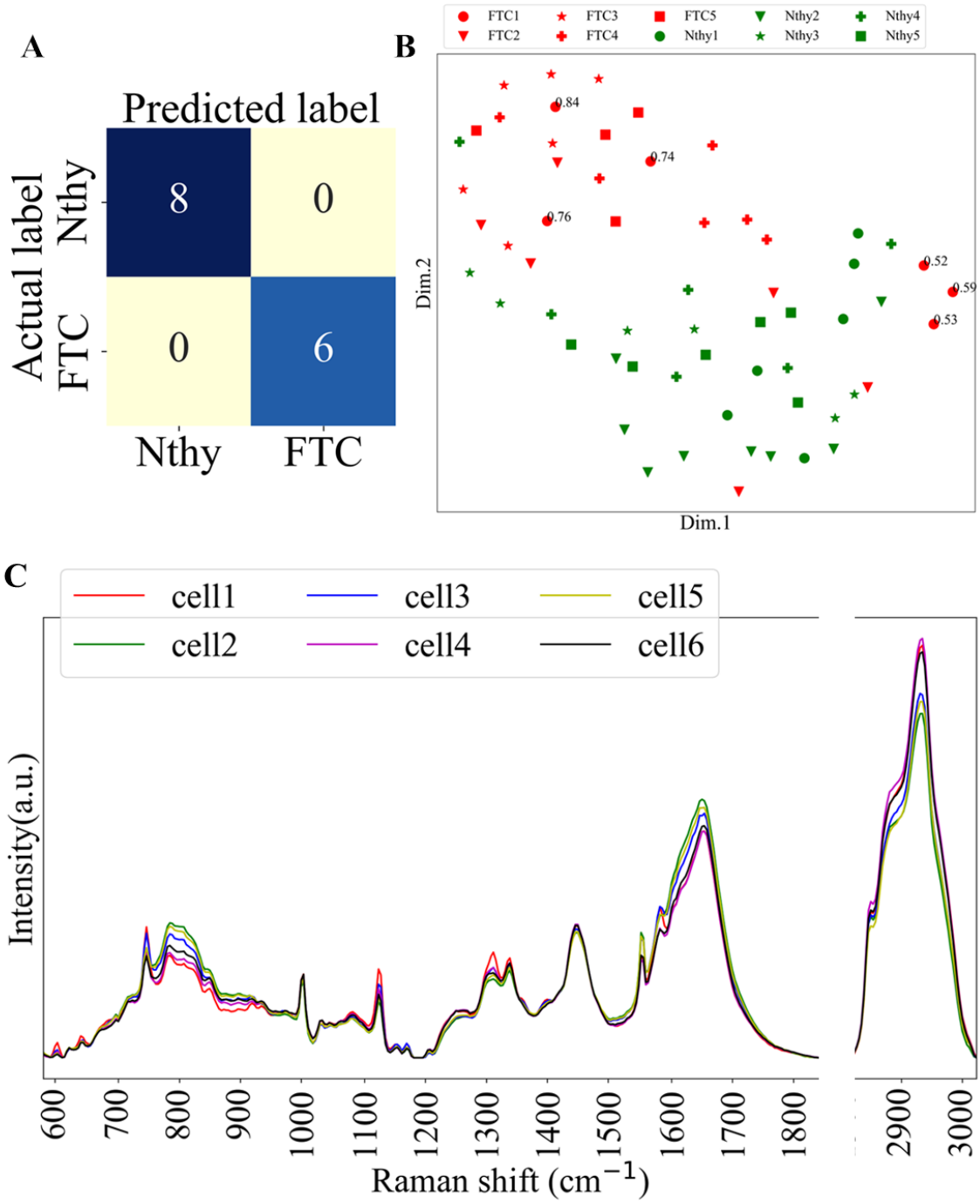


Fig. 4.23. For standard preprocessing with position-dependent wavenumber calibration: (A) Confusion matrix of test set (Nthy2 and FTC1) (B) predicted probability of 6 cells of FTC1 image in UMAP projection (C) average spectrum of 6 cells of FTC1 image. Note that remaining 8 images (4 Nthy and 4 FTC) were train set.

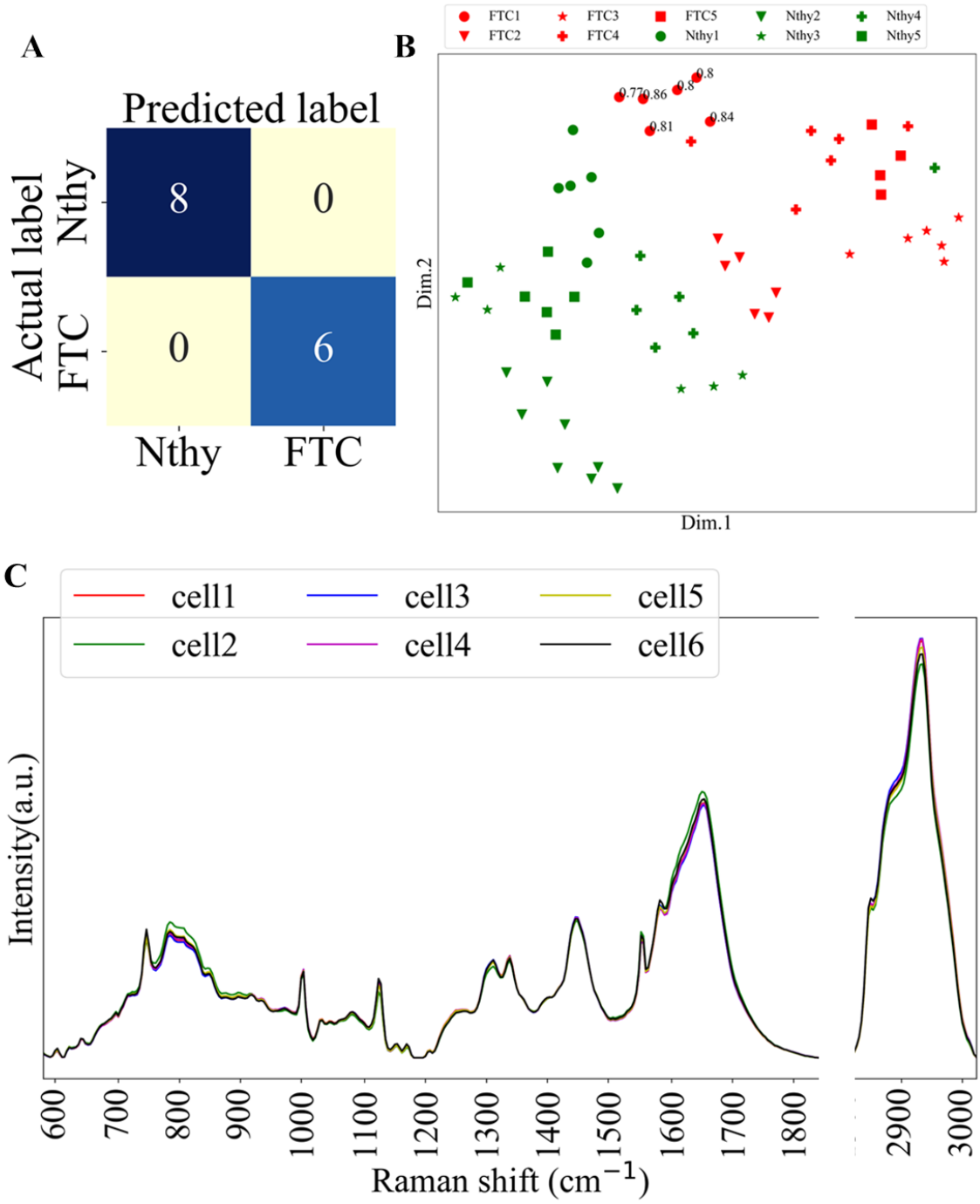


Fig. 4.24. For the detrending scheme after the implementation of position-dependent wavenumber calibration: (A) Confusion matrix of test set (Nthy2 and FTC1) (B) predicted probability of 6 cells of FTC1 image in UMAP projection (C) average spectrum of 6 cells of FTC1 image. Note that remaining 8 images (4 Nthy and 4 FTC) were train set.

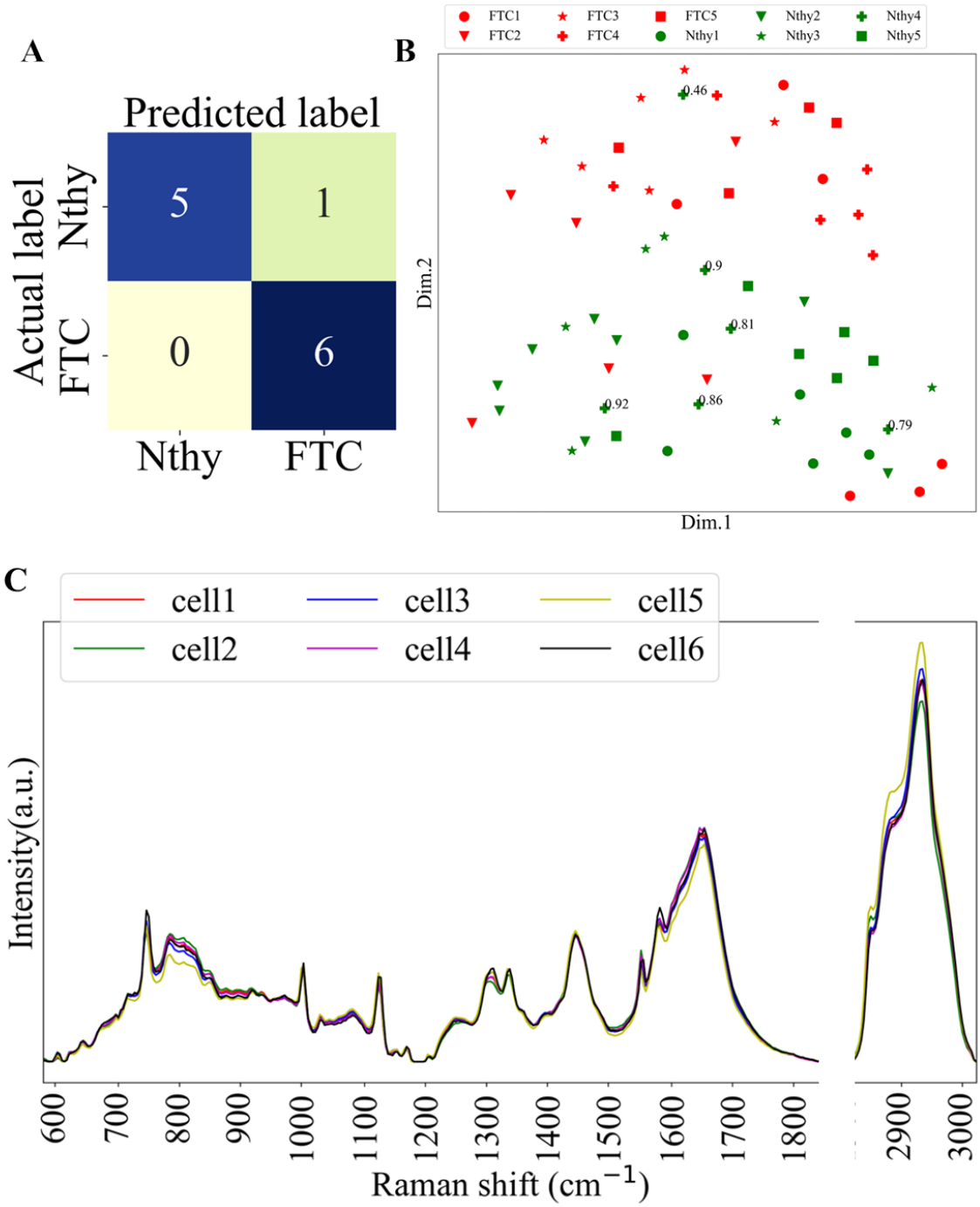


Fig. 4.25. For standard preprocessing without wavenumber calibration: (A) Confusion matrix of test set (Nthy4 and FTC2) (B) predicted probability of 6 cells of Nth4 image in UMAP projection (C) average spectrum of 6 cells of Nthy4 image. Note that remaining 8 images (4 Nthy and 4 FTC) were train set.

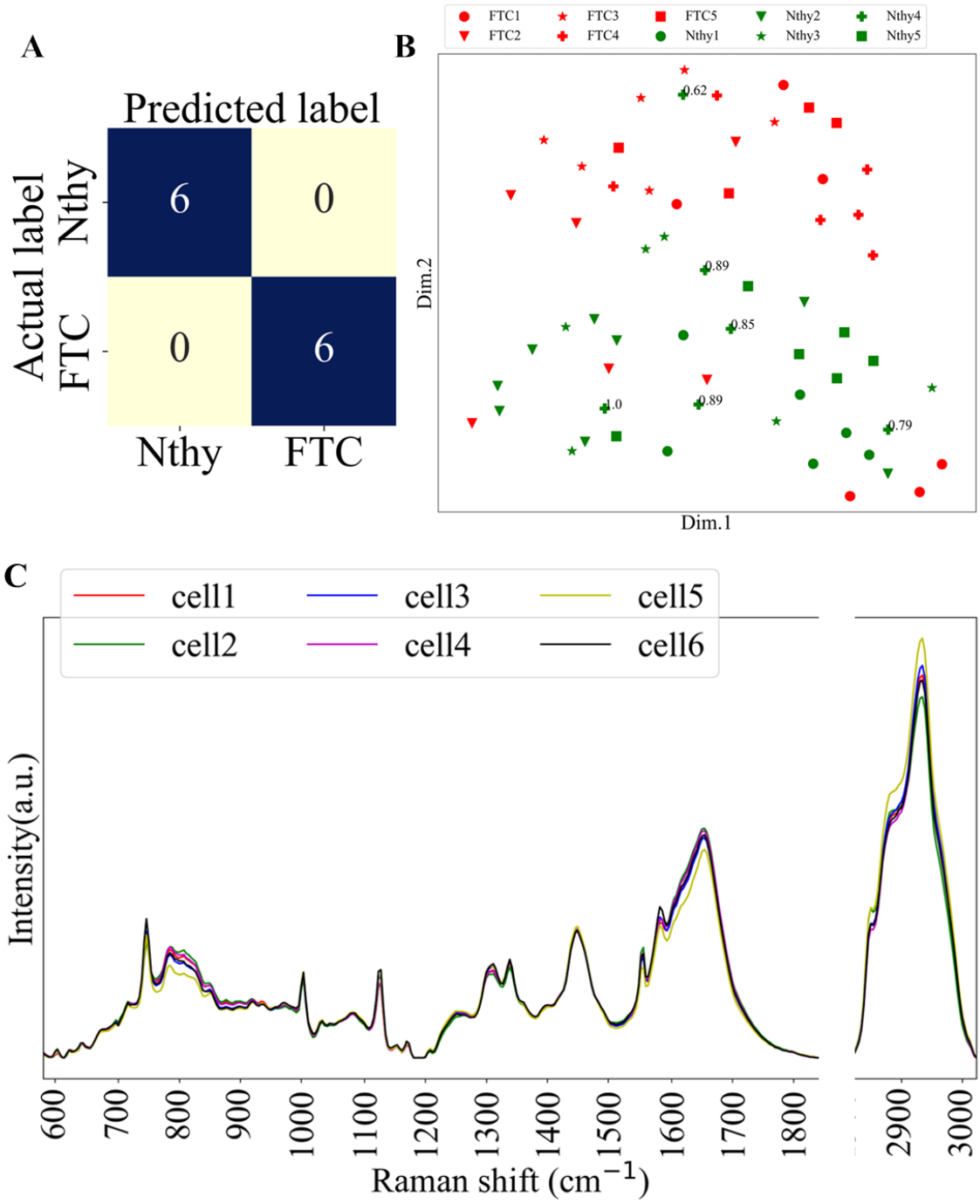


Fig. 4.26. For standard preprocessing with position-dependent wavenumber calibration: (A) Confusion matrix of test set (Nthy4 and FTC2) (B) predicted probability of 6 cells of Nth4 image in UMAP projection (C) average spectrum of 6 cells of Nthy4 image. Note that remaining 8 images (4 Nthy and 4 FTC) were train set.

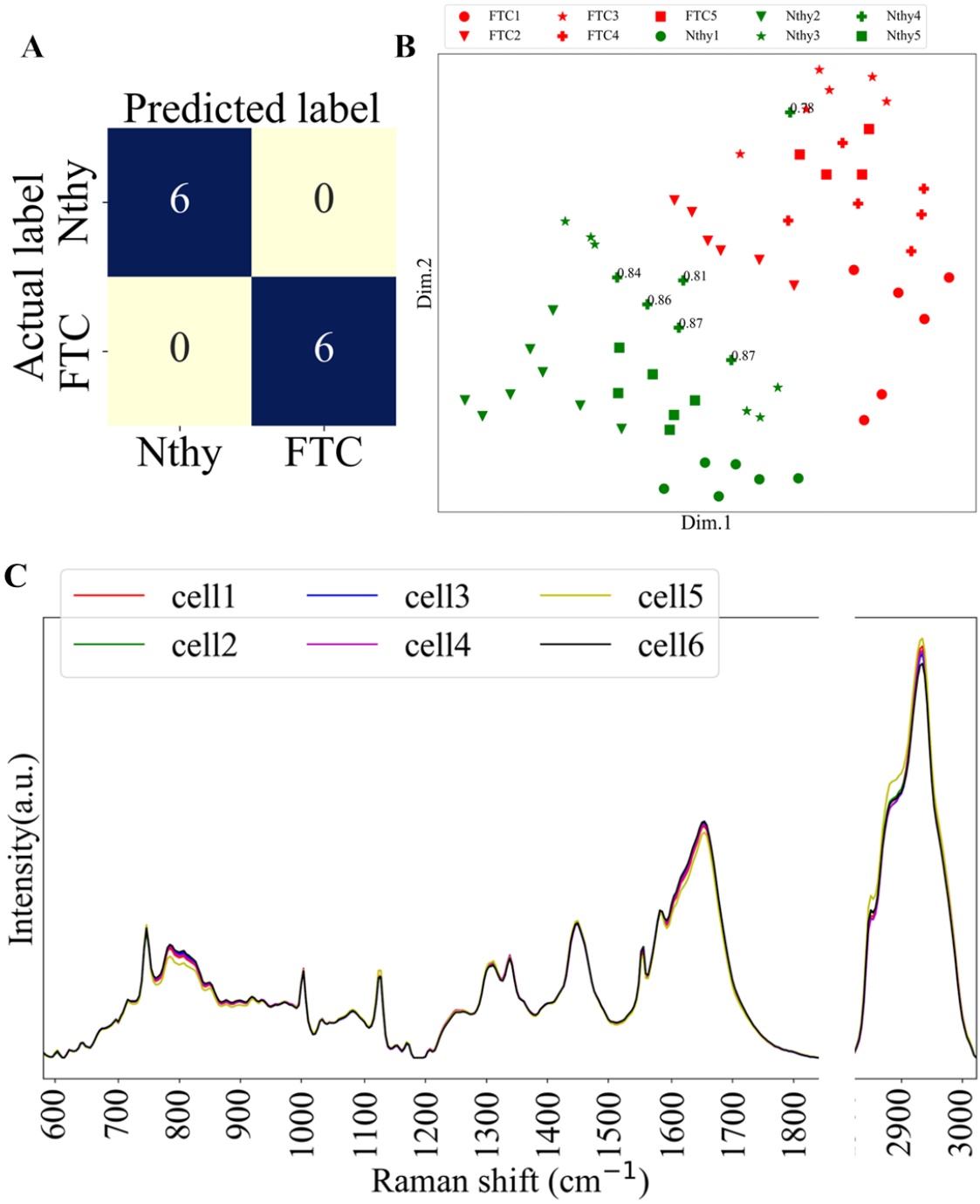


Fig. 4.27. For the detrending scheme after the implementation of position-dependent wavenumber calibration: (A) Confusion matrix of test set (Nthy4 and FTC2) (B) predicted probability of 6 cells of Nth4 image in UMAP projection (C) average spectrum of 6 cells of Nthy4 image. Note that remaining 8 images (4 Nthy and 4 FTC) were train set.

4.5 Conclusion

Preprocessing is a central aspect of microscopic data science pipeline, as it minimizes unwanted variations in data and enhances differentiability between different phenotypes assuming the underlying information can support it. To improve the standardization of hyperspectral Raman images acquired with line-scanning set-ups, we incorporated corrections in the spatial domain. In particular, we showed potential wavenumber drifts along the line illumination axis that altered the quality of the preprocessed Raman images. It has been shown that neglecting to consider a wavenumber calibration that varies with the pixel position along the illumination axis results in an apparent artificial spatial positive, negative, or small gradient in Raman images dependent on wavenumbers. Additionally we proved that standard preprocessing methods used in the field are ineffective in removing the influence of the non-homogeneous illumination, including slow-varying intensity fluctuations, in Raman images. Using standard preprocessing methods that do not correct spatial variations usually reduces the accuracy of the analysis and leads to misclassification of cells or questionable spectral composition of cells.

To address this issues we introduced a novel position-dependent wavenumber calibration to reflect the possible chromatic aberration or changes in physical properties resulting from laser light intensity variation along the illumination line direction, combined with a detrending scheme of spatial correlation along illumination and scanning directions, based on Karhunen-Loeve basis and a random forest regression. By using this proposed preprocessing strategy, enhanced differentiability was observed between phenotypes in the MDS plot and UMAP space, compared to the position-dependent wavenumber calibration. It should also be noted that, after the position-dependent wavenumber calibration, some (negligibly small) scars along the illumination axis were remained in the reconstructed Raman image, which was removed by random forest regression-based detrending protocol.

The remaining issue is the validation of our working hypothesis that, for Raman images, the individual PC score is not correlated to the physical space. We interpret that, if the sample distribution would actually have some apparent bias or trend in the physical space in the data set of Raman images, this hypothesis does not necessarily hold. Thus,

in actual applications, we must take into account how sufficiently the position-dependent wavenumber calibration eliminates artificial spatial biases, and how samples are distributed in the physical space over the data set to be analyzed, with a comparison of differentiability of phenotypes in the reconstructed Raman signals.

5

Conclusions and future plans

In this dissertation, the following important issues were emphasized:

In Chapter 1, general introduction and literature review regarding this current work were discussed.

In Chapter 2, some algorithms were discussed with illustrations that applied in this work.

In Chapter 3, I explained the three preprocessing schemes, namely uncalibrated, calibrated, and detrended schemes. In the uncalibrated scheme, I found some artificial shift of the peak position at some Raman shifts that create intensity variation among the cells in a Raman image along the line illumination axis and scanning axis as well. Depending on the image, some cell spectra are very different from others cell spectra. One can differentiate the spectra even though they come from the same image by the uncalibrated preprocessing scheme. However, peaks are aligned after calibrated scheme and intensity variation is now reduced

regardless the position of the cells. But still we have some intensity variation based on the position of the cells where they are located. After applying my detrended scheme, intensity distributions are now homogeneous irrespective to the position of the cells. Moreover, the spectra are closer to each other, and the variance of the spectra was reduced compared to the other two schemes.

In Chapter 4, detailed quantitative analyses of three preprocessing schemes spectra were explained. At first, I employed the Pearson correlation coefficients (r) between the Raman image at each individual wavenumber and the illumination axis ζ and the scanning axis ξ. A high value of the r was appeared for the uncalibrated preprocessing scheme which manifest that the high spatial dependency between Raman shift and spatial coordinates. After calibration, these dependency were reduced but after my detrending scheme nearly no dependency exists. Secondly, low dimensional projections (PCA,MDS,UMAP) shows very clear separation between two phenotypes by my detrending scheme compared to uncalibrated and calibrated schemes. Thirdly, accuracy, AUC, and f1 score were evaluated based on 25 cross-validation approaches. It is observed detrending scheme shows a high average with low variance in the value of these three performance measures. Fourthly, k-means clustering shows the homogeneous clusters within the cell by our detrending scheme compared to the other two schemes. Moreover, the predicted probability of some spectra shows a very marginal score near 0.5 while correct prediction by uncalibrated or calibrated schemes but our detrended scheme shows the high score and closer scores for all spectra.

The feature selection[85–87] also called feature/variable importance describes which features are relevant. It helps us a better understanding of the data and sometimes assists model improvements by employing the feature selection. If a dataset has thousand of features then probably some features may be redundant, some of the features may be correlated and some of the features may be irrelevant for the model. If we use all the features, it will require a huge amount of time to train the model, and model performance will be reduced. So, feature selection is vitally important in model building. The purpose of feature selection is to build the best probable model without redundant and irrelevant features. On the other hand, Class Activation Mapping (CAM) was first introduced by Zhou

et al. [88] in their paper titled “Learning Deep Features for Discriminative Localization" to identify the regions of an input image that are most relevant to the network’s prediction for a particular class. Grad-CAM (Gradient-weighted Class Activation Mapping) is a variant of CAM, was introduced by Selvaraju et al. [89] that provides a more detailed and better visualization of the important regions in an image. In the future, I am planning to employ some feature selection methods, including Random Forest, Partial Least Squares (PLS) regression variable importance scores, and ANOVA, to identify important features/variables. Additionally, I intend to utilize CAM and Grad-CAM techniques to identify important regions within Raman spectra.

Appendices

A

Supporting Information

A.1 Parameters selection for SVD and baseline correction

The choice of 8 components has been determined based on the maximization of the classification accuracy between FTC-133 and Nthy-ori 3-1, and the minimization of the signal distortion. The parameters for both SVD denoising and polynomial fitting choices were chosen based on the optimization of 25-fold cross validation accuracy for various pair of hyperparameters. Fig. A.1 shows the accuracy distribution dependency on the order of SVD and polynomial fitting for baseline correction. As seen in Fig. A.1, the polynomial order 6 among the polynomial orders tested (6, 8 and 10) consistently resulted in the largest standard deviations. The pairing of SVD denoising by keeping 8 components with a 8^{th} order polynomial model, denoted here by [SVD:Polyfit]=[8:8], resulted in the smallest standard deviation. Although the mean accuracy of [8:10] is slightly higher than [8:8], the standard deviation of accuracy is larger for [8:10] than for [8:8]. Thus, we chose [SVD:Polyfit]=[8:8] in this work. To further clarify the signal distortion caused by retaining only a few SVD components in the denoising phase, we refer to Fig. A.2. When retaining only 4 SVD components, the intensity profiles of cytochrome, protein and lipid wavenumbers tend to be similar meaning we filtered out some important chemical information from Raman spectra. On the other hand, when retaining anywhere from 8 SVD components to 20 components, the difference in shape among cytochrome, protein and lipids distribution becomes noticeable. Moreover the shape of the distribution remains stable over the range of components used.

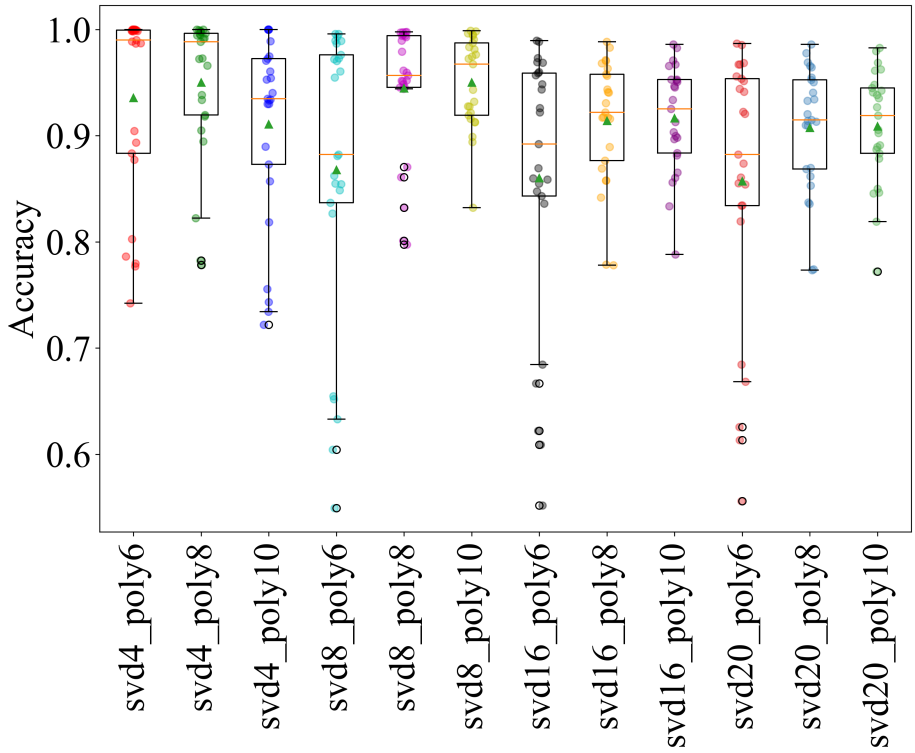


Fig. A.1. The box-and-whisker plot of test accuracy in the prediction of FTC-133/Nthy-ori 3-1 for the standard preprocessing without wavenumber calibration of 25-fold cross validation based on pixelwise spectra: different pairs of singular value decomposition components for denoising and polynomial fitting orders for baseline corrections.

Samples used and their representative and average spectra before and after processing are shown in Fig. A.3.

At the pixel level spectra are noisy. For example, the figure Fig. A.4 shows the raw spectra and a denoised version. The noise is also detectable in the spatial domain as can be seen in a raw Raman image at wavenumber 1548 cm^{-1}. After applying SVD denoising, the irregularity of the signal is minimized.

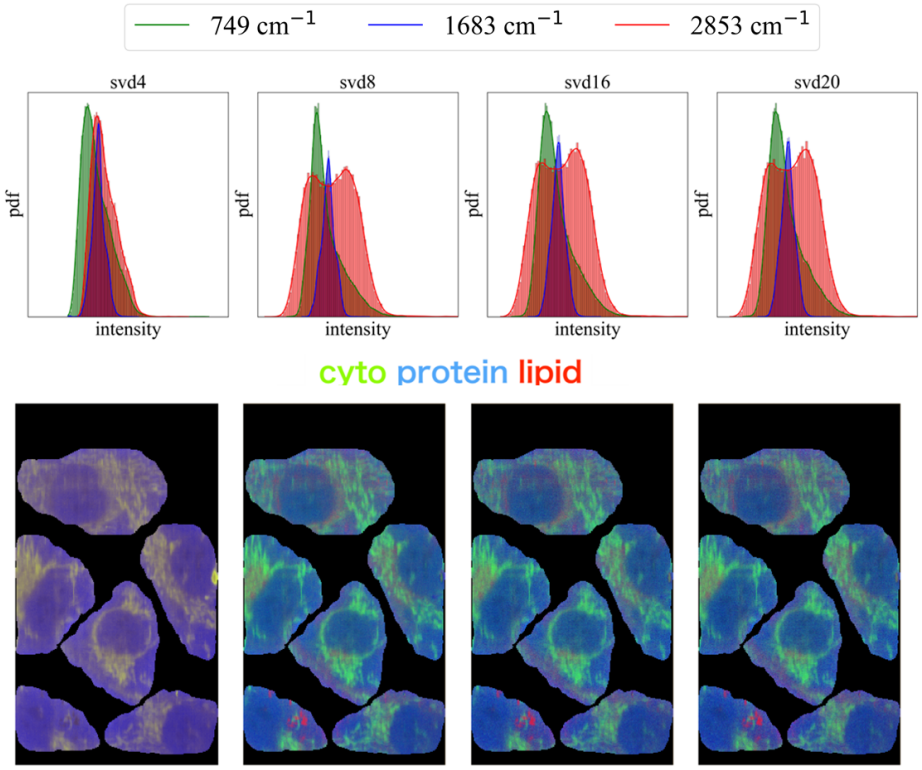


Fig. A.2. The Raman intensity distribution at known cytochrome peak 749 cm^{-1}, protein peak 1683 cm^{-1} and lipid peak 2853 cm^{-1} for the Raman image of FTC-133(#2) for different singular value decomposition components.

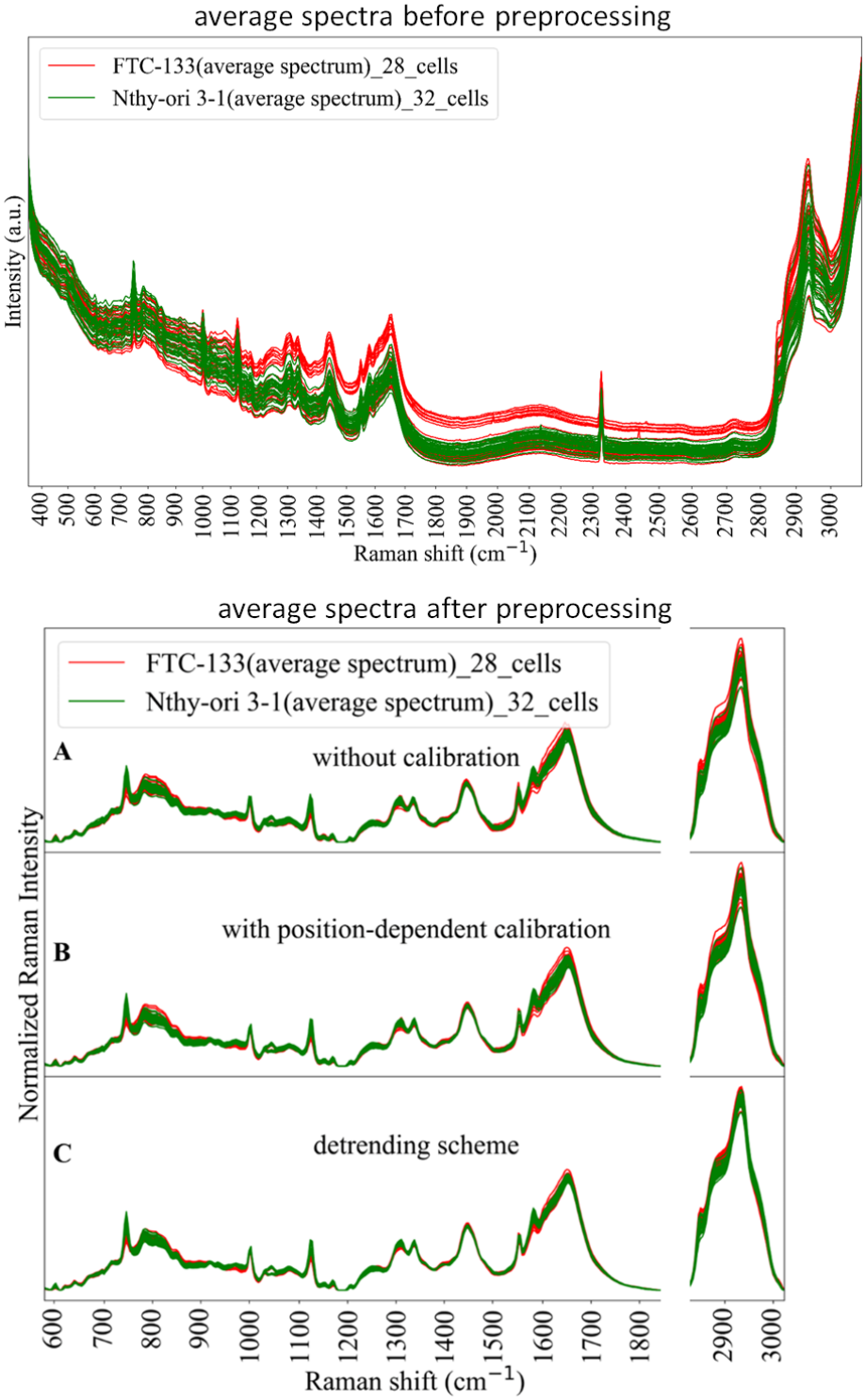


Fig. A.3. Average spectra of 32 Nthy-ori 3-1 and 28 FTC-133 cells: before preprocessing (top) (A-C) standard preprocessing without wavenumber calibration, standard preprocessing with position-dependent wavenumber calibration,the detrending scheme.

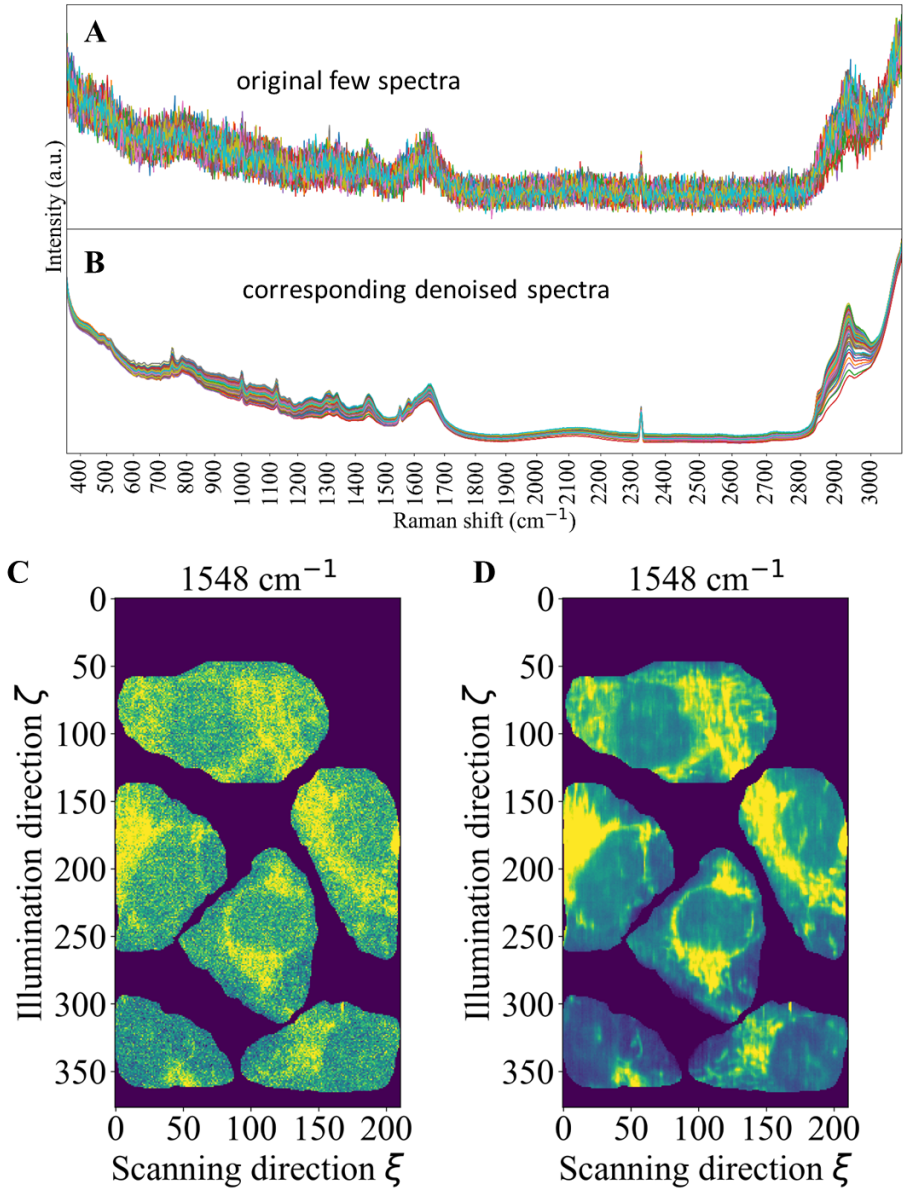


Fig. A.4. For the Raman image of FTC-133(#2) (A) few individual pixel level raw Raman spectra (B) corresponding denoised spectra, (C-D) intensity distribution at 1548 cm^{-1}: (C) raw Raman image, (D) denoised Raman image.

A.2 Advantages of the random forest model with other models

To demonstrate the advantages of the random forest model, we showed the comparison with a series of polynomial regressions of different order 3 to 12 as well as the averaged PC score as an alternative way to estimate the unwanted trend in a Raman image. Specifically, to estimate the trend for each individual PC score, we averaged the set of points of PC scores at each position of the illumination axis along the scanning axis. After subtracting the trend along the illumination axis of each PC, we estimate the trend for each PC along the scanning axis by taking the average along the illumination axis. To compare the effects of these trend estimation methods (random forest regression, polynomial regressions and average PC score), the set of preprocessed data was then visualized in the low dimensional space using classical multidimensional scaling (MDS) (Fig. A.5). In this MDS space, we observed that the separability between Nthy-ori 3-1 and FTC-133 improved when the data was either corrected by the random forest procedure or average PC scores, compared to polynomial regression. Indeed, polynomial regressions models are global models that cannot describe some abrupt changes in the trend by such polynomial order of 3-12, as opposed to local strategies by random forest and average PC scores that handle these changes much more effectively. Fig. A.6 exemplifies the resultant Raman images at different wavenumbers 807, 1294, and 1407 cm^{-1} extracted by polynomial fitting of order 8, random forest regression (RFR), and averge PC score. We can see that some scars (indicated by orange arrow marks in the figure), materialized by stripes in Raman images, are present in the reconstructed Raman images processed by polynomial regression, but are minimized when corrected by random forest or local average PC scores. Furthermore, the classification accuracy results, performed with the 25-fold cross-validation shows that random forest has the higher performance, with a higher average accuracy associated with a lower standard deviation accuracy (shown in Fig. A.7). We also observed that detrending by averaged PC scores gives similar performance to RF regression. In this example, although the performance of RF regression and average PC score schemes show a similar accuracy in the classification between FTC-133 and Nthy-ori 3-1, we privilege random forest (RF) regression over average PC scores. The reason is that the average PC score scheme removes any trend that exists in Raman images along illumination and scanning directions by definition. On the contrary, RF regression corresponds to a series of step functions whose step width can be adjusted by a hyperparameter, and also takes into account statistics generated by bootstrap sampling using a limited data set (In Chapter2, a brief explanation Random Forest regression was given). The default value of the hyperparameter we used in the paper corresponded to a step width being one pixel. However, in theory, there should also exist some samples whose spatial trend along illumination and/or scanning axis may not necessarily arise from optics but from sample's nature themselves. In such situation, there exists

a room in random forest regression that the hyperparameter can be tuned for specific needs in estimating the trend in diverse experimental contexts.

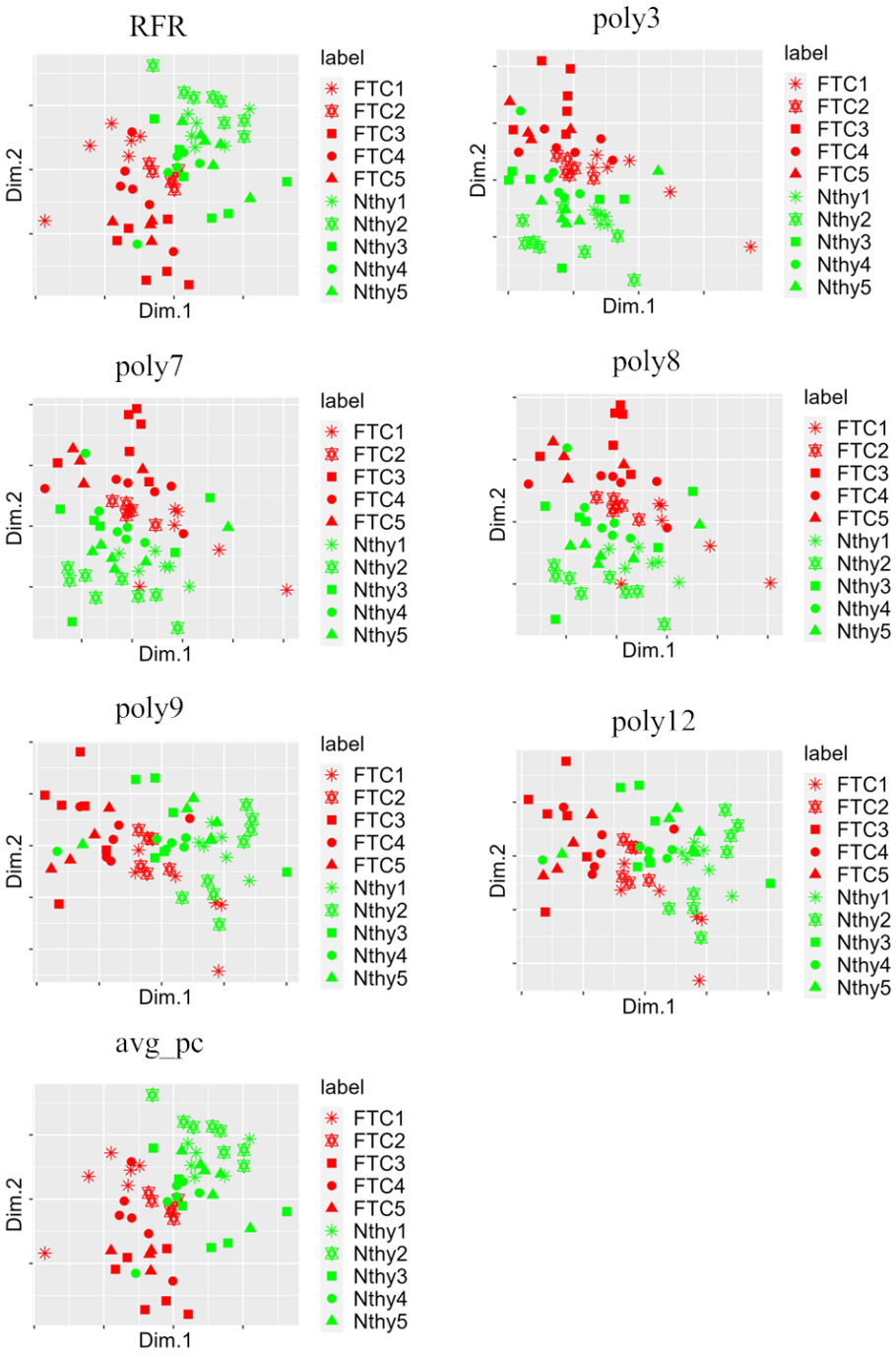


Fig. A.5. The multi-dimensional scaling (MDS) projection of the ten Raman images including sixty single cells in total for seven detrending methods.

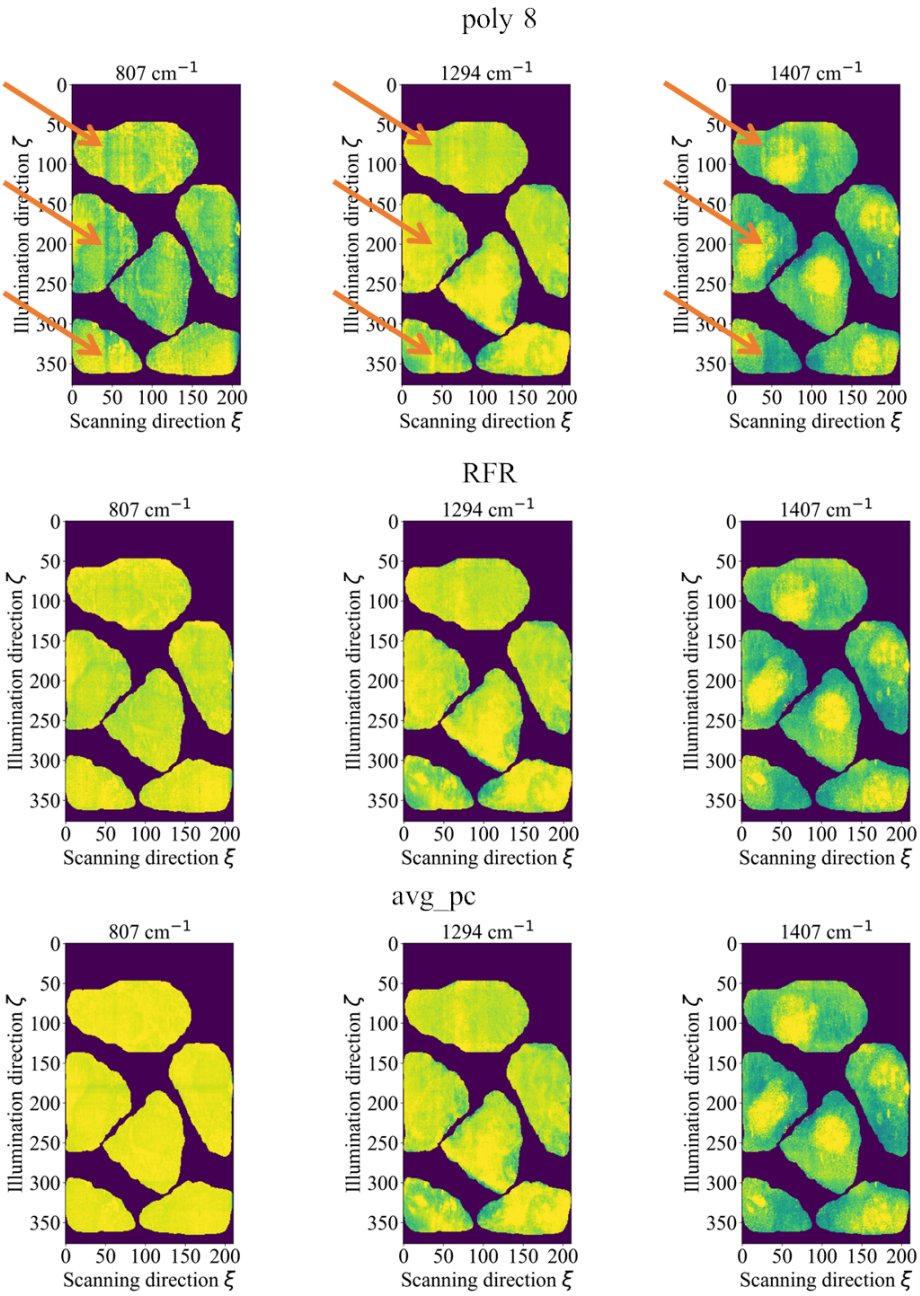


Fig. A.6. The Raman intensity distribution at three different peaks 807 cm^{-1}, 1294 cm^{-1}, 1407 cm^{-1} in the space domain for the Raman image of FTC-133(#2): polynomial fitting of order 8 (top row), Random forest regression (middle row), average PC (bottom row). Orange arrows indicate some fictitious straight lines (scars) created by polynomial fitting scheme.

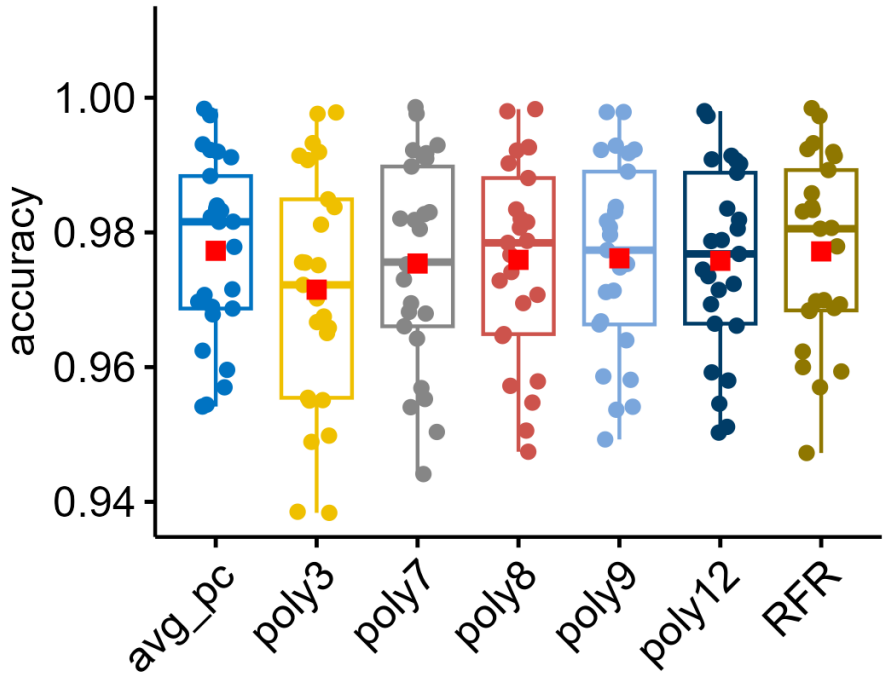


Fig. A.7. The box-and-whisker plot of test accuracy in the prediction of FTC-133/Nthy-ori 3-1 for seven detrending methods of 25-fold cross validation based on pixelwise spectra.

A.3 Measurement of Dimethyl sulfoxide (DMSO)

To evaluate the workflow on a homogeneous substrate, we measured a Raman image of Dimethyl sulfoxide (DMSO) with a line exposure time of 3s. We added Fig. A.8 and Fig. A.9 to demonstrate that our detrending scheme corrects the non-homogeneous profile. As observed for cell samples, Fig. A.8 tells us that the “uncalibrated" data set demonstrates a high correlation between the Raman shift and the spatial coordinates. The calibrated data had a lower correlation, whereas, as expected, the data corrected by our detrending workflow did not demonstrate any correlations. As seen in Fig. A.8 that the Raman image at 2912 cm^{-1} has a homogeneous intensity distribution compared to the non-corrected ones. One can also see in Fig. A.8F that after area-normalization all three schemes (with/without position-dependent wave number calibration and detrending method) provide almost indistinguishable Raman spectra of DMSO. Thus, the issue is that without detrending scheme non-homogeneous profile remains in practice. To further confirm whether our detrending scheme can recover the chemical homogeneity of the DMSO sample, more effectively than either the standard (uncalibrated data) and/or the position dependent wavenumber calibration scheme (calibrated data), a k-means clustering with $k = 3$ was performed independently on the sets of Raman spectra preprocessed with the three different schemes. The cluster assignment for each spectrum of the DMSO image is reported Fig. A.9A-C. We observe that the uncalibrated and the calibrated preprocessed Raman spectra present a gradient of group assignment. This indicates that the intensity variation along the illumination axis corrupt the expected chemical homogeneity. Conversely, the k-means cluster map estimated after using the detrending scheme demonstrates a random mixing of the three clusters, suggesting that we are able to restore a degree of a chemical homogeneity throughout the entire space. As seen in Fig. A.9D, in the principal component (PC) space constructed with all preprocessed Raman images i.e., (position-dependently) noncalibrated, (position-dependently) calibrated, and detrended, we found the detrended Raman spectra are projected onto a very localized region. In contrast, noncalibrated and calibrated spectra tend to spread over a larger area of the PC space. This observation further supports that our detrending scheme effectively reduces intensity variability throughout a Raman image.

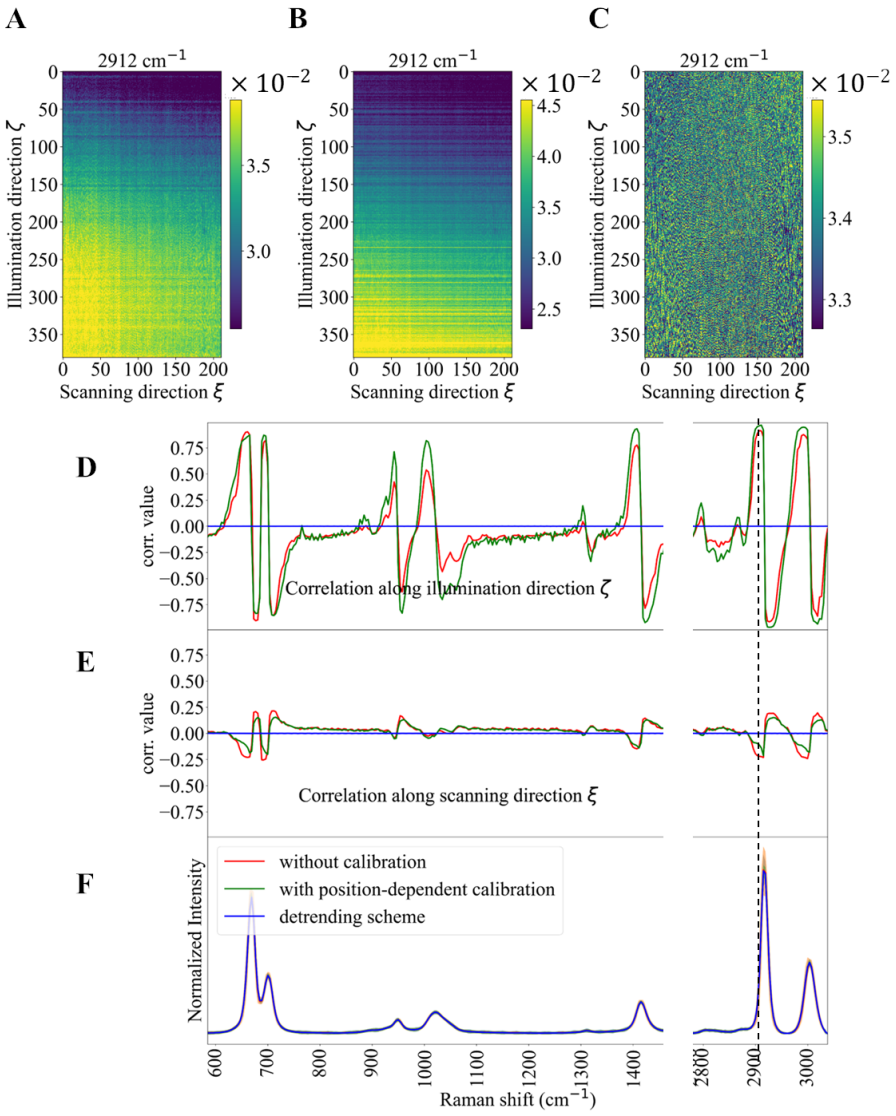


Fig. A.8. (A)-(C) The Raman intensity distribution at 2912 cm^{-1} (dashed vertical line) in the space domain of DMSO: (A) after standard preprocessing without (position-dependent) wavenumber calibration, (B) after standard preprocessing with position-dependent wavenumber calibration, (C) after the detrending scheme applied on the top of position-dependent wavenumber calibration. (D)-(E) The Pearson correlation coefficients between the Raman images at each wavenumber acquired by the three preprocessings: (D) the illumination axis coordinate, (E) the scanning axis coordinate. (F) The average with two standard deviation Raman spectra over whole regions obtained by the three different schemes.

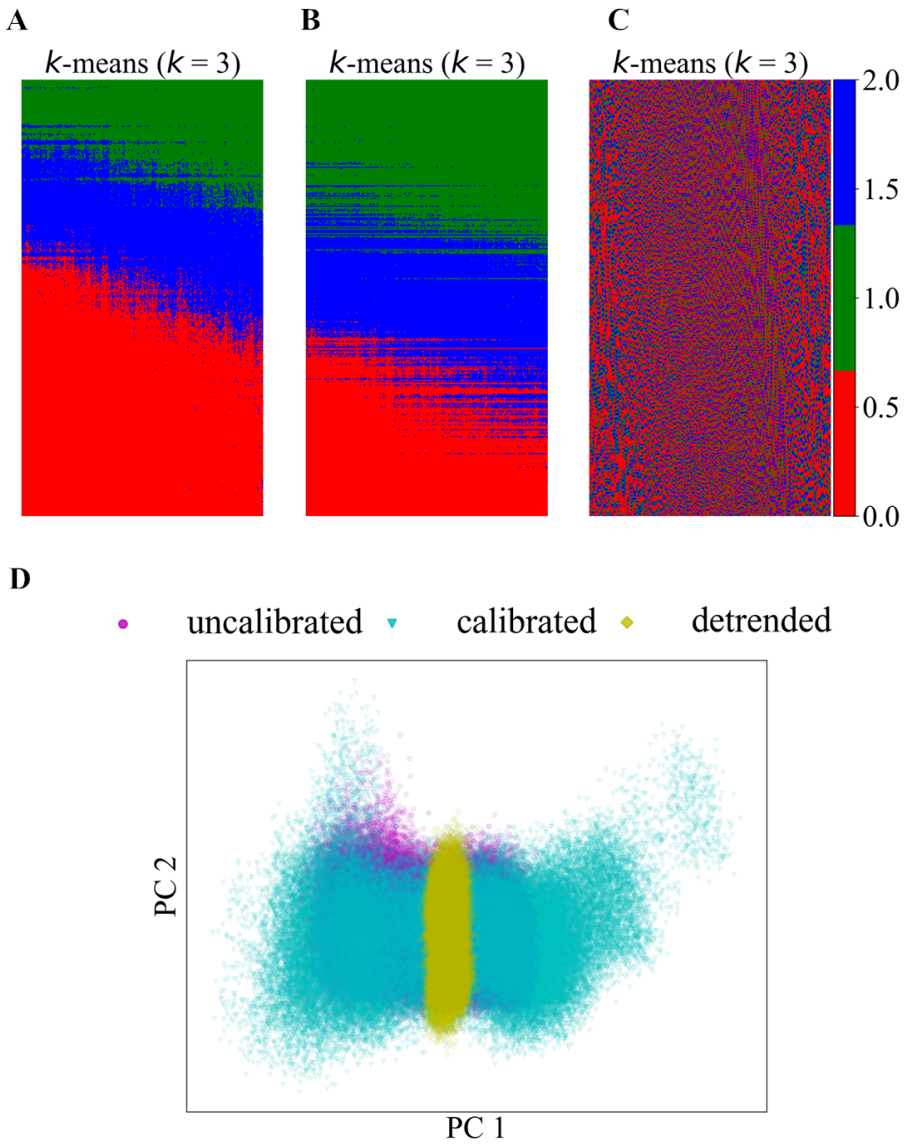


Fig. A.9. (A)-(C) The k-means clustering maps with $k = 3$ for individual Raman spectra in the Raman image for DMSO: (A) standard preprocessing without position-dependent wavenumber calibration (B) the position-dependent wavenumber calibration (C) the detrending scheme. (D) PCA projection of all spectra based on three preprocessing schemes.

References

[1] Thomas Bocklitz, Angela Walter, Katharina Hartmann, Petra Rösch, and Jürgen Popp. How to pre-process raman spectra for reliable and stable models? *Analytica chimica acta*, 704(1-2):47–56, 2011.

[2] J Nicholas Taylor, Kentaro Mochizuki, Kosuke Hashimoto, Yasuaki Kumamoto, Yoshinori Harada, Katsumasa Fujita, and Tamiki Komatsuzaki. High-resolution raman microscopic detection of follicular thyroid cancer cells with unsupervised machine learning. *The Journal of Physical Chemistry B*, 123(20):4358–4372, 2019.

[3] Robert Nakayama, Keisuke Horiuchi, Michiro Susa, Seiichi Hosaka, Yuichiro Hayashi, Kaori Kameyama, Yoshihisa Suzuki, Hiroo Yabe, Yoshiaki Toyama, and Hideo Morioka. Anaplastic transformation of follicular thyroid carcinoma in a metastatic skeletal lesion presenting with paraneoplastic leukocytosis. *Thyroid*, 22(2):200–204, 2012.

[4] Zanyar Movasaghi, Shazza Rehman, and Ihtesham U Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007.

[5] Rajinder Singh. Cv raman and the discovery of the raman effect. *Physics in Perspective*, 4(4):399–420, 2002.

[6] Huilu Yao, Zhanhua Tao, Min Ai, Lixin Peng, Guiwen Wang, Bijuan He, and Yongqing Li. Raman spectroscopic analysis of apoptosis of single human gastric cancer cells. *Vibrational Spectroscopy*, 50(2):193–197, 2009.

[7] Giuseppe Pezzotti. Raman spectroscopy in cell biology and microbiology. *Journal of Raman Spectroscopy*, 52(12):2348–2443, 2021.

[8] Alessandra D'Avanzo, Patrick Treseler, Philip HG Ituarte, Mariwil Wong, Leanne Streja, Francis S Greenspan, Allan E Siperstein, Quan-Yang Duh, and Orlo H Clark. Follicular thyroid carcinoma: histology and prognosis. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 100(6):1123–1129, 2004.

[9] Christopher R McHenry and Roy Phitayakorn. Follicular adenoma and carcinoma of the thyroid gland. *The oncologist*, 16(5):585, 2011.

[10] Manuel Sobrinho-Simoes, Catarina Eloy, Joao Magalhaes, Cláudia Lobo, and Teresina Amaro. Follicular thyroid carcinoma. *Modern Pathology*, 24(2):S10–S18, 2011.

[11] Robert A Robinson. *Head and neck pathology: atlas for histologic and cytologic diagnosis*. Lippincott Williams & Wilkins, 2012.

[12] Giorgio Grani, Livia Lamartina, Cosimo Durante, Sebastiano Filetti, and David S Cooper. Follicular thyroid cancer and hürthle cell carcinoma: challenges in diagnosis, treatment, and clinical management. *The Lancet Diabetes & Endocrinology*, 6(6):500–514, 2018.

[13] Arno Germond, Taro Ichimura, Takaaki Horinouchi, Hideaki Fujita, Chikara Furusawa, and Tomonobu M Watanabe. Raman spectral signature reflects transcriptomic features of antibiotic resistance in escherichia coli. *Communications biology*, 1(1):1–10, 2018.

[14] Iwan W Schie and Thomas Huser. Methods and applications of raman microspectroscopy to single-cell analysis. *Applied spectroscopy*, 67(8):813–828, 2013.

[15] Pablo Manuel Ramos and Itziar Ruisánchez. Noise and background removal in raman spectra of ancient pigments using wavelet transform. *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering*, 36(9):848–856, 2005.

[16] Todd Hollon and Daniel A Orringer. Label-free brain tumor imaging using raman-based methods. *Journal of Neuro-Oncology*, 151(3):393–402, 2021.

[17] Matthew E Martin, Musundi B Wabuyele, Kui Chen, Paul Kasili, Masoud Panjehpour, Mary Phan, Bergein Overholt, Glenn Cunningham, Dale Wilson, Robert C DeNovo, et al. Development of an advanced hyperspectral imaging (hsi) system with applications for cancer detection. *Annals of biomedical engineering*, 34(6):1061–1068, 2006.

[18] Almar F Palonpon, Mikiko Sodeoka, and Katsumasa Fujita. Molecular imaging of live cells by raman microscopy. *Current opinion in chemical biology*, 17(4):708–715, 2013.

[19] Ujjwal Karn. An intuitive explanation of convolutional neural networks. *The data science blog*, 2016.

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Tushar Jajodia and Pankaj Garg. Image classification–cat and dog images. *Image*, 6(23):570–572, 2019.

[24] Shefali Arora and MP S Bhatia. Handwriting recognition using deep learning in keras. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 142–145. IEEE, 2018.

[25] Yian Seo and Kyung-shik Shin. Hierarchical convolutional neural networks for fashion image classification. *Expert systems with applications*, 116:328–339, 2019.

[26] Georgy V Konoplich, Evgeniy O Putin, and Andrey A Filchenkov. Application of deep learning to the problem of vehicle detection in uav images. In *2016 XIX IEEE International Conference on Soft Computing and Measurements (SCM)*, pages 4–6. IEEE, 2016.

[27] Jiabin Xia, Lianqing Zhu, Mingxin Yu, Tao Zhang, Zhihui Zhu, Xiaoping Lou, Guangkai Sun, and Mingli Dong. Analysis and classification of oral tongue squamous cell carcinoma based on raman spectroscopy and convolutional neural networks. *Journal of Modern Optics*, 67(6):481–489, 2020.

[28] Jinchao Liu, Margarita Osadchy, Lorna Ashton, Michael Foster, Christopher J Solomon, and Stuart J Gibson. Deep convolutional neural networks for raman spectrum recognition: a unified solution. *Analyst*, 142(21):4067–4074, 2017.

[29] Chi-Sing Ho, Neal Jean, Catherine A Hogan, Lena Blackmon, Stefanie S Jeffrey, Mark Holodniy, Niaz Banaei, Amr AE Saleh, Stefano Ermon, and Jennifer Dionne. Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning. *Nature communications*, 10(1):1–8, 2019.

[30] Wooje Lee, Aufried TM Lenferink, Cees Otto, and Herman L Offerhaus. Classifying raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *Journal of raman spectroscopy*, 51(2):293–300, 2020.

[31] Xuejing Chen, Luyuan Xie, Yonghong He, Tian Guan, Xuesi Zhou, Bei Wang, Guangxia Feng, Haihong Yu, and Yanhong Ji. Fast and accurate decoding of raman spectra-encoded suspension arrays using deep learning. *Analyst*, 144(14):4312–4319, 2019.

[32] Xiaqiong Fan, Wen Ming, Huitao Zeng, Zhimin Zhang, and Hongmei Lu. Deep learning-based component identification for the raman spectra of mixtures. *Analyst*, 144(5):1789–1798, 2019.

[33] Ugur Parlatan, Medine Tuna Inanc, Bahar Yuksel Ozgor, Engin Oral, Ercan Bastu, Mehmet Burcin Unlu, and Gunay Basar. Raman spectroscopy as a non-invasive diagnostic technique for endometriosis. *Scientific reports*, 9(1):1–7, 2019.

[34] Michael Jermyn, Kelvin Mok, Jeanne Mercier, Joannie Desroches, Julien Pichette, Karl Saint-Arnaud, Liane Bernstein, Marie-Christine Guiot, Kevin Petrecca, and Frederic Leblond. Intraoperative brain cancer detection with raman spectroscopy in humans. *Science translational medicine*, 7(274):274ra19–274ra19, 2015.

[35] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6):e0177544, 2017.

[36] Danying Ma, Linwei Shang, Jinlan Tang, Yilin Bao, Juanjuan Fu, and Jianhua Yin. Classifying breast cancer tissue by raman spectroscopy with one-dimensional convolutional neural network. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 256:119732, 2021.

[37] Jiaqi Hu, De Zhang, Hantao Zhao, Biao Sun, Pei Liang, Jiaming Ye, Zhi Yu, and Shangzhong Jin. Intelligent spectral algorithm for pigments visualization, classification and identification based on raman spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 250:119390, 2021.

[38] Gang Yin, Lintao Li, Shun Lu, Yu Yin, Yuanzhang Su, Yilan Zeng, Mei Luo, Maohua Ma, Hongyan Zhou, Lucia Orlandini, et al. An efficient primary screening of covid-19 by serum raman spectroscopy. *Journal of Raman Spectroscopy*, 2021.

[39] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[40] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

[41] Haifeng Cao, Wei Ji, Zhuang Liu, and Yanglin Zhou. Principal component analysis. 1987.

[42] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[43] Kirk Baker. Singular value decomposition tutorial. *The Ohio State University*, 24, 2005.

[44] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[45] ER Henry and J Hofrichter. [8] singular value decomposition: Application to analysis of experimental data. In *Methods in enzymology*, volume 210, pages 129–192. Elsevier, 1992.

[46] Michael C Hout, Megan H Papesh, and Stephen D Goldinger. Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):93–103, 2013.

[47] Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of computational and graphical statistics*, 17(2):444–472, 2008.

[48] Hervé Abdi. Metric multidimensional scaling (mds): analyzing distance matrices. *Encyclopedia of measurement and statistics*, pages 1–13, 2007.

[49] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[50] Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015.

[51] Tim Hesterberg. Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):497–526, 2011.

[52] V Rodriguez-Galiano, M Sanchez-Castillo, M Chica-Olmo, and MJOGR Chica-Rivas. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818, 2015.

[53] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1):1–16, 2009.

[54] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, et al. Using raman spectroscopy to characterize biological materials. *Nature protocols*, 11(4):664–687, 2016.

[55] Chad A Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological raman spectra. *Applied spectroscopy*, 57(11):1363–1367, 2003.

[56] John R Ferraro. *Introductory raman spectroscopy*. Elsevier, 2003.

[57] Richard L McCreery. *Raman spectroscopy for chemical analysis*. John Wiley & Sons, 2005.

[58] Imran I Patel and Francis L Martin. Discrimination of zone-specific spectral signatures in normal human prostate using raman spectroscopy. *Analyst*, 135(12):3060–3069, 2010.

[59] Khalifa Mohammad Helal, James Nicholas Taylor, Harsono Cahyadi, Akira Okajima, Koji Tabata, Yoshito Itoh, Hideo Tanaka, Katsumasa Fujita, Yoshinori Harada, and Tamiki Komatsuzaki. Raman spectroscopic histology using machine learning for nonalcoholic fatty liver disease. *FEBS letters*, 593(18):2535–2544, 2019.

[60] Tsunehisa Yamamoto, Takeo Minamikawa, Yoshinori Harada, Yoshihisa Yamaoka, Hideo Tanaka, Hitoshi Yaku, and Tetsuro Takamatsu. Label-free evaluation of myocardial infarct in surgically excised ventricular myocardium by raman spectroscopy. *Scientific Reports*, 8(1):14671, 2018.

[61] José Luis González-Solís, Juan Carlos Martínez-Espinosa, Juan Manuel Salgado-Román, and Pascual Palomares-Anda. Monitoring of chemotherapy leukemia treatment using raman spectroscopy and principal component analysis. *Lasers in medical science*, 29:1241–1249, 2014.

[62] Menglu Li, Hao-Xiang Liao, Kazuki Bando, Yasunori Nawa, Satoshi Fujita, and Katsumasa Fujita. Label-free monitoring of drug-induced cytotoxicity and its molecular fingerprint by live-cell raman and autofluorescence imaging. *Analytical Chemistry*, 94(28):10019–10026, 2022.

[63] Shreyas Rangan, H Georg Schulze, Martha Z Vardaki, Michael W Blades, James M Piret, and Robin FB Turner. Applications of raman spectroscopy in the development of cell therapies: State of the art and future perspectives. *Analyst*, 145(6):2070–2105, 2020.

[64] Cordelia Orillac, Todd Hollon, and Daniel A Orringer. Clinical translation of stimulated raman histology. *Biomedical Engineering Technologies: Volume 1*, pages 225–236, 2022.

[65] Yanping Li, Binglin Shen, Shaowei Li, Yihua Zhao, Junle Qu, and Liwei Liu. Review of stimulated raman scattering microscopy techniques and applications in the biosciences. *Advanced Biology*, 5(1):2000184, 2021.

[66] Yasuaki Kumamoto, Yoshinori Harada, Tetsuro Takamatsu, and Hideo Tanaka. Label-free molecular imaging and analysis by raman spectroscopy. *Acta histochemica et cytochemica*, 51(3):101–110, 2018.

[67] Jason D Rodriguez, Benjamin J Westenberger, Lucinda F Buhse, and John F Kauffman. Standardization of raman spectra for transfer of spectral libraries across different instruments. *Analyst*, 136(20):4232–4240, 2011.

[68] Jasper Engel, Jan Gerretzen, Ewa Szymańska, Jeroen J Jansen, Gerard Downey, Lionel Blanchet, and Lutgarde MC Buydens. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, 2013.

[69] Agnieszka Martyna, Alicja Menżyk, Alessandro Damin, Aleksandra Michalska, Gianmario Martra, Eugenio Alladio, and Grzegorz Zadora. Improving discrimination of raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components. *Chemometrics and Intelligent Laboratory Systems*, 202:104029, 2020.

[70] Nils Kristian Afseth, Vegard Herman Segtnan, and Jens Petter Wold. Raman spectra of biological samples: A study of preprocessing methods. *Applied spectroscopy*, 60(12):1358–1367, 2006.

[71] Peter Lasch. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems*, 117:100–114, 2012.

[72] Shuxia Guo, Jürgen Popp, and Thomas Bocklitz. Chemometric analysis in raman spectroscopy from experimental design to machine learning–based modeling. *Nature Protocols*, 16(12):5426–5459, 2021.

[73] Keisaku Hamada, Katsumasa Fujita, Nicholas Isaac Smith, Minoru Kobayashi, Yasushi Inouye, and Satoshi Kawata. Raman microscopy for dynamic molecular imaging of living cells. *Journal of biomedical optics*, 13(4):044027–044027, 2008.

[74] Ji Qi and Wei-Chuan Shih. Performance of line-scan raman microscopy for high-throughput chemical imaging of cell population. *Applied Optics*, 53(13):2881–2885, 2014.

[75] Hao He, Mengxi Xu, Cheng Zong, Peng Zheng, Lilan Luo, Lei Wang, and Bin Ren. Speeding up the line-scan raman imaging of living cells by deep convolutional neural network. *Analytical chemistry*, 91(11):7070–7077, 2019.

[76] Michael D Graham and Ioannis G Kevrekidis. Alternative approaches to the karhunen-loeve decomposition for model reduction and data analysis. *Computers & chemical engineering*, 20(5):495–506, 1996.

[77] Thomas Dörfer, Thomas Bocklitz, Nicolae Tarcea, Michael Schmitt, and Jürgen Popp. Checking and improving calibration of raman spectra using chemometric approaches. *Zeitschrift für Physikalische Chemie*, 225(6-7):753–764, 2011.

[78] TW 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.

[79] Christian Darken and John Moody. Fast adaptive k-means clustering: some empirical results. In *1990 IJCNN international joint conference on neural networks*, pages 233–238. IEEE, 1990.

[80] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[81] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[82] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[83] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[84] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[85] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992.

[86] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[87] B Venkatesh and J Anuradha. A review of feature selection and its methods. *Cybernetics and information technologies*, 19(1):3–26, 2019.

[88] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[89] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[90] Jan Gerretzen, Ewa Szymańska, Jacob Bart, Antony N Davies, Henk-Jan van Manen, Edwin R van den Heuvel, Jeroen J Jansen, and Lutgarde MC Buydens. Boosting model performance and interpretation by entangling preprocessing selection and variable selection. *Analytica Chimica Acta*, 938:44–52, 2016.

[91] Nils Kristian Afseth and Achim Kohler. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 117:92–99, 2012.

[92] Shuxia Guo, Thomas Bocklitz, Ute Neugebauer, and Jürgen Popp. Common mistakes in cross-validating classification models. *Analytical Methods*, 9(30):4410–4417, 2017.

[93] Jan Gerretzen, Ewa Szymańska, Jeroen J Jansen, Jacob Bart, Henk-Jan van Manen, Edwin R van den Heuvel, and Lutgarde MC Buydens. Simple and effective way for data preprocessing selection based on design of experiments. *Analytical chemistry*, 87(24):12096–12103, 2015.

[94] Chang He, Shuo Zhu, Xiaorong Wu, Jiale Zhou, Yonghui Chen, Xiaohua Qian, and Jian Ye. Accurate tumor subtype detection with raman spectroscopy via variational autoencoder and machine learning. *ACS omega*, 7(12):10458–10468, 2022.

[95] Ryan S Jakubek and Marc D Fries. Calibration of the temporal drift in absolute and relative raman intensities in large raman images using a mercury–argon lamp. *Journal of Raman Spectroscopy*, 53(1):137–147, 2022.

[96] Luke R Sadergaski, Travis J Hager, and Hunter B Andrews. Design of experiments, chemometrics, and raman spectroscopy for the quantification of hydroxylammonium, nitrate, and nitric acid. *ACS omega*, 7(8):7287–7296, 2022.

[97] Jeon Woong Kang, Yun Sang Park, Hojun Chang, Woochang Lee, Surya Pratap Singh, Wonjun Choi, Luis H Galindo, Ramachandra R Dasari, Sung Hyun Nam, Jongae Park, et al. Direct observation of glucose fingerprint using in vivo raman spectroscopy. *Science Advances*, 6(4):eaay5206, 2020.

[98] Bianca Durrant, Matthew Trappett, Dustin Shipp, and Ioan Notingher. Recent developments in spontaneous raman imaging of living biological cells. *Current opinion in chemical biology*, 51:138–145, 2019.

[99] Ian Khaw, Benjamin Croop, Jialei Tang, Anna Möhl, Ulrike Fuchs, and Kyu Young Han. Flat-field illumination for quantitative fluorescence imaging. *Optics express*, 26(12):15276–15288, 2018.

[100] Vipin Kumar and Sonajharia Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[101] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

[102] K Krishna and M Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999.

List of Publications and Presentations

Journal

1. Abdul Halim Bhuiyan, Jean-Emmanuel Cĺement, Zannatul Ferdous, Kentaro Mochizuki, Koji Tabata, James Nicholas Taylor, Yasuaki Kumamoto, Yoshinori Harada, Thomas Bocklitz, Katsumasa Fujita, and Tamiki Komatsuzaki, "Differentiability of cell types enhanced by detrending non-homogeneous pattern in line-illumination Raman microscope", *Analyst, 2023, 148, 3574–3583, DOI: 10.1039/d3an00516j*.

Conference Abstracts/Posters/Presentations

1. Abdul Halim Bhuiyan, Jean-Emmanuel Cĺement, Kentaro Mochizuki, James Nicholas Taylor, Koji Tabata, Yuta Mizuno, Atsuyoshi Nakamura, Yoshinori Harada, Katsumasa Fujita, and Tamiki Komatsuzaki, To classify Raman spectra using Deep Learning Approach, *21st RIES-HOKUDAI International Symposium*, Hokkaido University, Japan, December 10-11, 2020.

2. Abdul Halim Bhuiyan, Jean-Emmanuel Cĺement, Kentaro Mochizuki, James Nicholas Taylor, Koji Tabata, Yuta Mizuno, Atsuyoshi Nakamura, Yoshinori Harada, Katsumasa Fujita, and Tamiki Komatsuzaki, Feature extraction and classification of Raman spectra using Convolutional Neural Network Approach, *The 12th CSE International Autumn School and The 9th ALP International Symposium*, Hokkaido University, Japan, October 13-14, 2021.

3. Abdul Halim Bhuiyan, Jean-Emmanuel Cĺement, Kentaro Mochizuki, James Nicholas Taylor, Koji Tabata, Yuta Mizuno, Atsuyoshi Nakamura, Yoshinori Harada, Katsumasa Fujita, and Tamiki Komatsuzaki, To diagnose Follicular Thyroid Carcinoma using Convolutional Neural Network Approach, *22nd RIES-HOKUDAI International Symposium*, Hokkaido University, Japan, December 6-7, 2021.

4. Abdul Halim Bhuiyan, Jean-Emmanuel Cĺement, Kentaro Mochizuki, James Nicholas Taylor, Koji Tabata, Yuta Mizuno, Atsuyoshi Nakamura, Yoshinori Harada, Kat-

sumasa Fujita, and Tamiki Komatsuzaki, Understanding the results of black box Convolution Neural Network to identify Follicular thyroid cancer, *The 60th Annual Meeting of the Biophysical Society of Japan*, Hakodate Arena, Hokkaido, Japan, September 28 - 30, 2022.