Supplementary Information:

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

1 Data preprocessing

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 correction: 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[1]. 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. (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)[2] of order 8. 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. We newly added Fig. S19, which shows the accuracy distribution dependency on the order of SVD and polynomial fitting for baseline correction. As seen in Fig. S19, 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 an eighth-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. S20. 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. (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}).

2 Detrending scheme

After preprocessing (i.e., standard preprocessing with position-dependent wavenumber calibration) a Raman image $\hat{u}$ of size (m, n, ν), its 2D unfolded (=preprocessed) version $\hat{u}_{\text{unfold}}$ of size (nm, ν) is expanded in an orthonormal basis known as Karhunen-Loève (K-L) basis or principal component (PC) basis. In this expansion, the original spectral matrix, $\hat{u}_{\text{unfold}}$ which is mean-centered is expressed as the multiplication of two matrices : T of size (nm, ν) and V of size (ν, ν), such that $\hat{u}_{\text{unfold}} = TV^t$ with the transpose matrix of V, V^t. The matrix T represents the new coordinates of the Raman spectra, named as principal component (PC) scores, in a set of orthonormal basis vectors represented as the column vectors of V. Each ith PC score is translated into a 2D map of $n \times m$ pixels, denoted as $Q_i(k, l)$ with $1 \leq k \leq n$ and $1 \leq l \leq m$.

The detrended PC maps Raman images along both the illumination and scanning axes, denoted by $\bar{u}_{\text{unfold}}$, is finally provided by $\bar{u}_{\text{unfold}} = \bar{T}V^t$, where $\bar{T}$ is the detrended PC score matrix of size (nm, ν) whose columns are constructed

from $\bar{Q}_i$ instead of Q_i. Then we refold in 3D format the object $\bar{u}_{\text{unfold}} + M$ where M is a matrix of size (nm, ν) filled for each row with the average spectrum of all pixels of the original Raman image $\hat{u}$.

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 Figs. S14 and S15 to demonstrate that our detrending scheme corrects the non-homogeneous profile. As observed for cell samples, Fig. S14 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. S14 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. S14F 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. S15A-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. S15D, 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.

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 along with your suggestion. 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. S16). 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. S17 exemplifies the resultant Raman images at different wavenumbers 807, 1294, ans 1407 cm^{-1} extracted by polynomial fitting of order 8, random forest regression (RFR), and average 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. S18). 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 (See also Section 4). 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.

3 Average Raman spectra variation of individual cells in terms of the three preprocessings

Fig. S11 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. S11G and S11H 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.

4 Random Forest regression

We explain the procedure of random forest regression by using a simple illustrative example in Fig. S23A 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 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. S23A, 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. S24, 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. S24A, 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[3–5]. 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. S23B 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 paper, we 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.

References

[1] M. E. Wall, A. Rechtsteiner and L. M. Rocha, in *A Practical Approach to Microarray Data Analysis*, Springer, 2003, pp. 91–109.

[2] C. A. Lieber and A. Mahadevan-Jansen, *Applied Spectroscopy*, 2003, **57**, 1363–1367.

[3] T. Hesterberg, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2011, **3**, 497–526.

Table S1: Assignment of the important peaks in the Raman spectra of FTC-133 and Nthy-ori 3-1 cell lines[6–8].

Raman peaks	
Wavenumber peak(±3cm^{-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

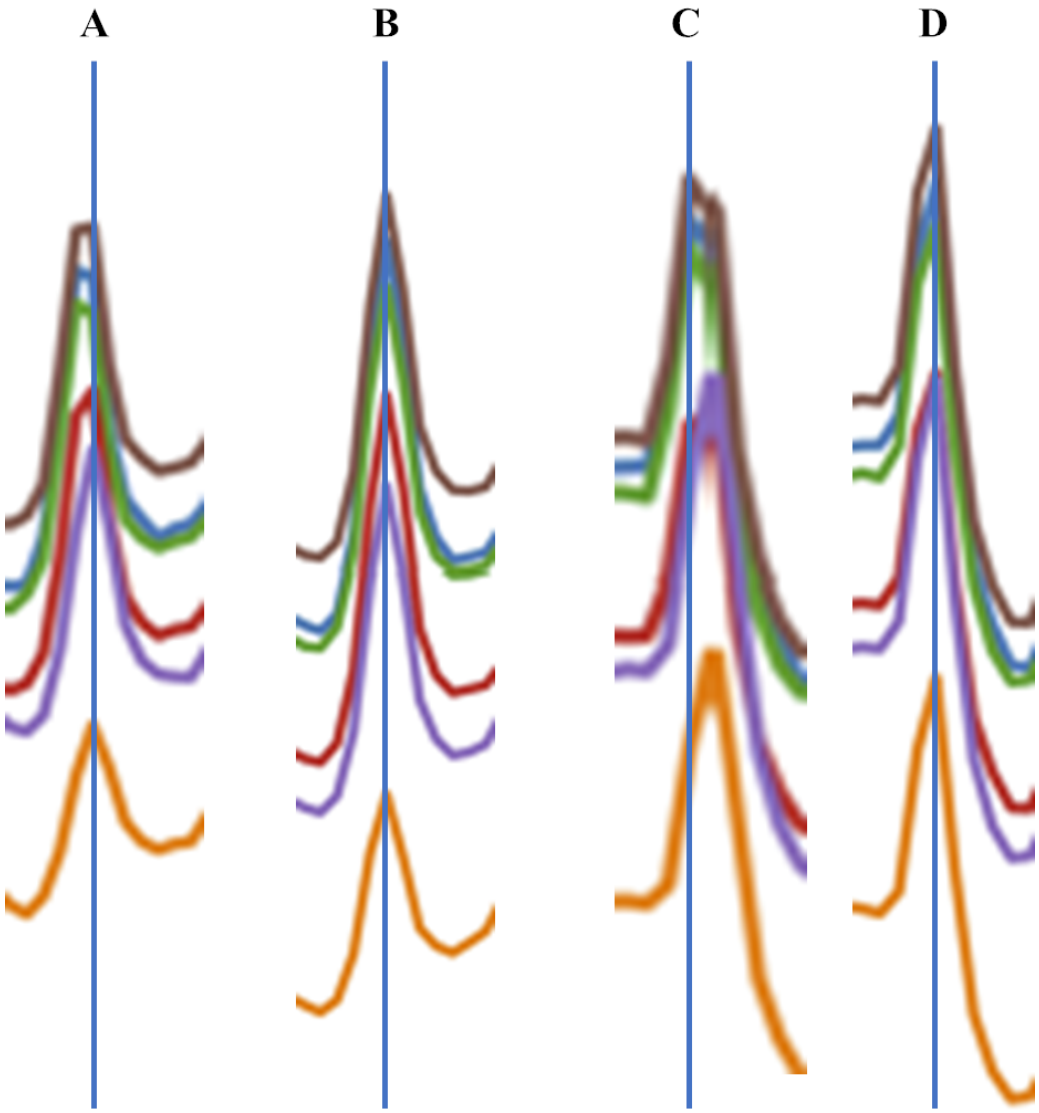

Figure S1: Enlarged peak positions of the averaged spectra of six cells contained in a Raman image of FTC-133(#2) at two Raman shifts (A and C) without wavenumber calibration, (B and D) with the position-dependent wavenumber calibration.

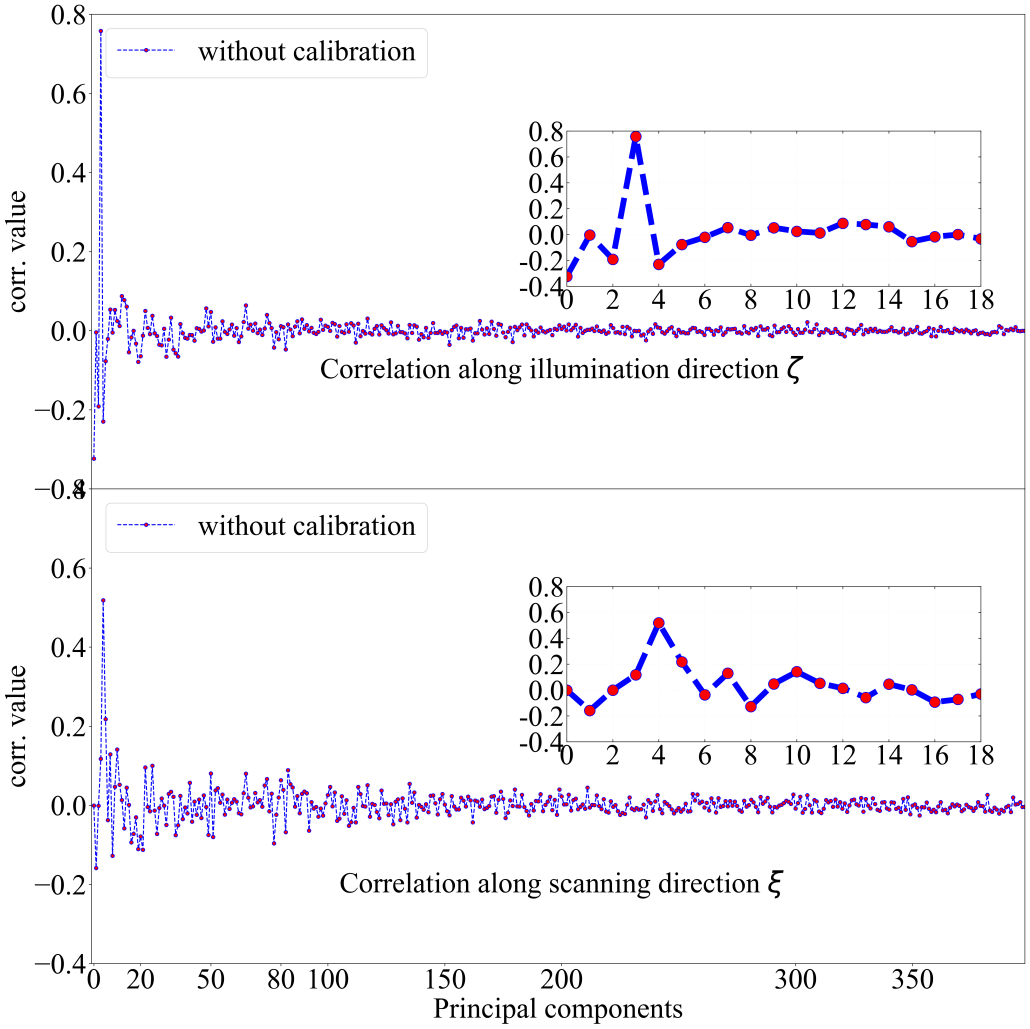

Figure S2: The Pearson correlation coefficients between the spatial coordinates, illumination and scanning axes, and the images of PCs of a Raman image of FTC-133(#2) preprocessed by standard preprocessing without wavenumber calibration.

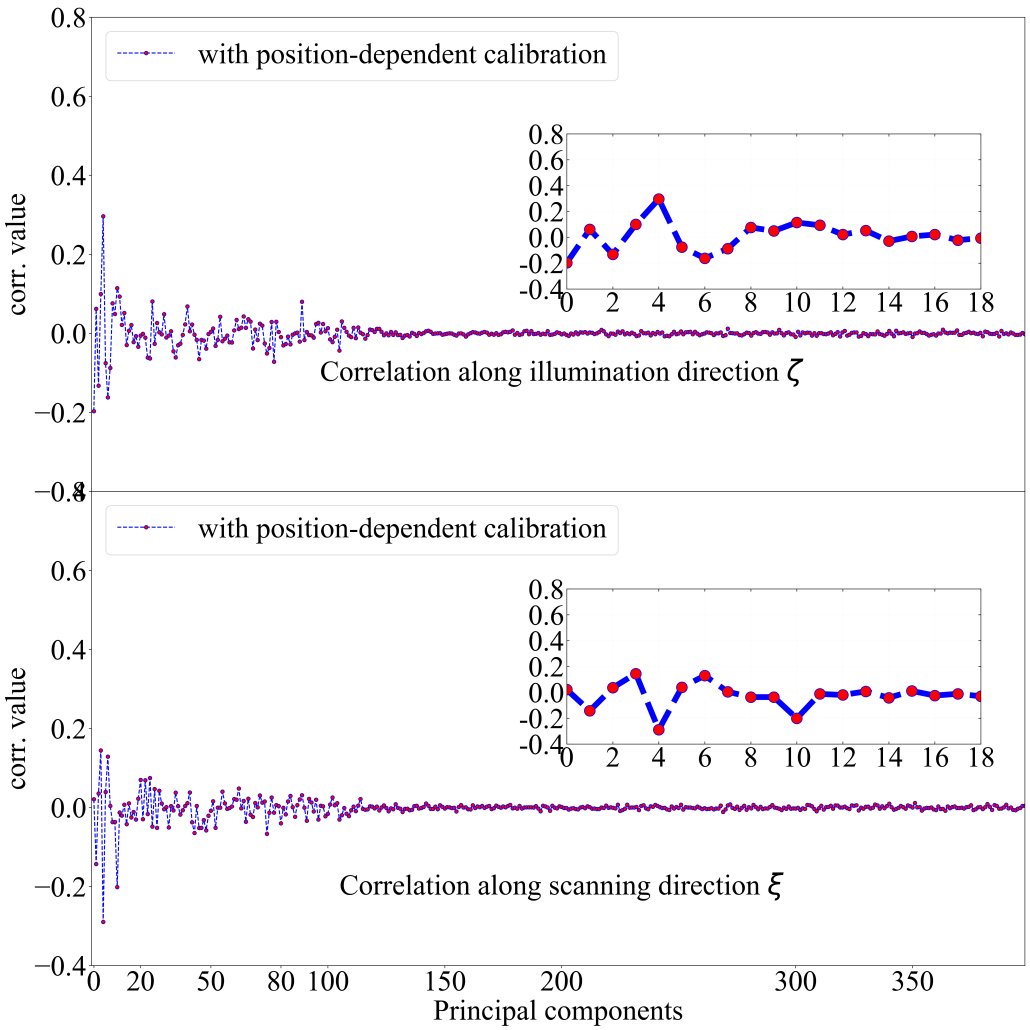

Figure S3: The Pearson correlation coefficients between the spatial coordinates, illumination and scanning axes, and the images of PCs a Raman image of FTC-133(#2) preprocessed by standard preprocessing with the position-dependent wavenumber calibration.

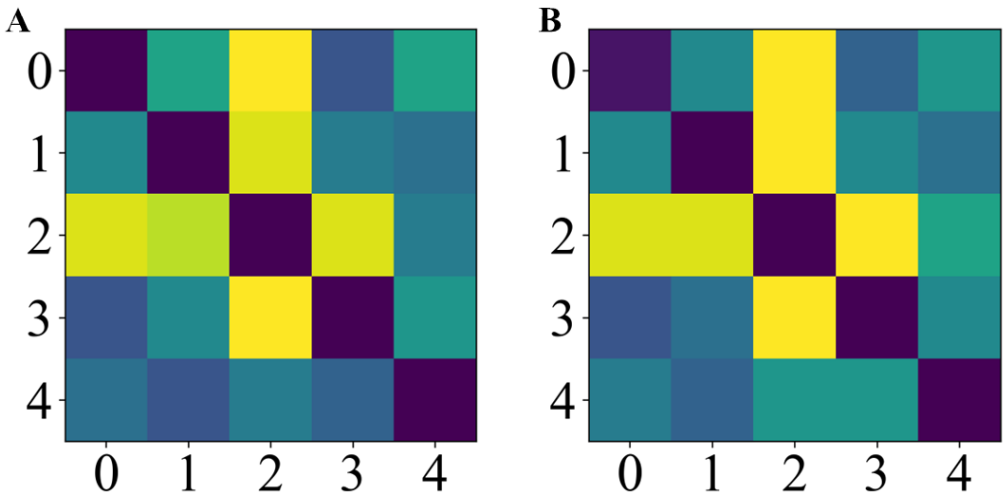

Figure S4: The distance matrix between the centroid of each cluster obtained for the Raman image of FTC-133(#2):(A) with the position-dependent wavenumber calibration vs without wavenumber calibration and (B) with the position-dependent wavenumber calibration vs the detrending scheme.

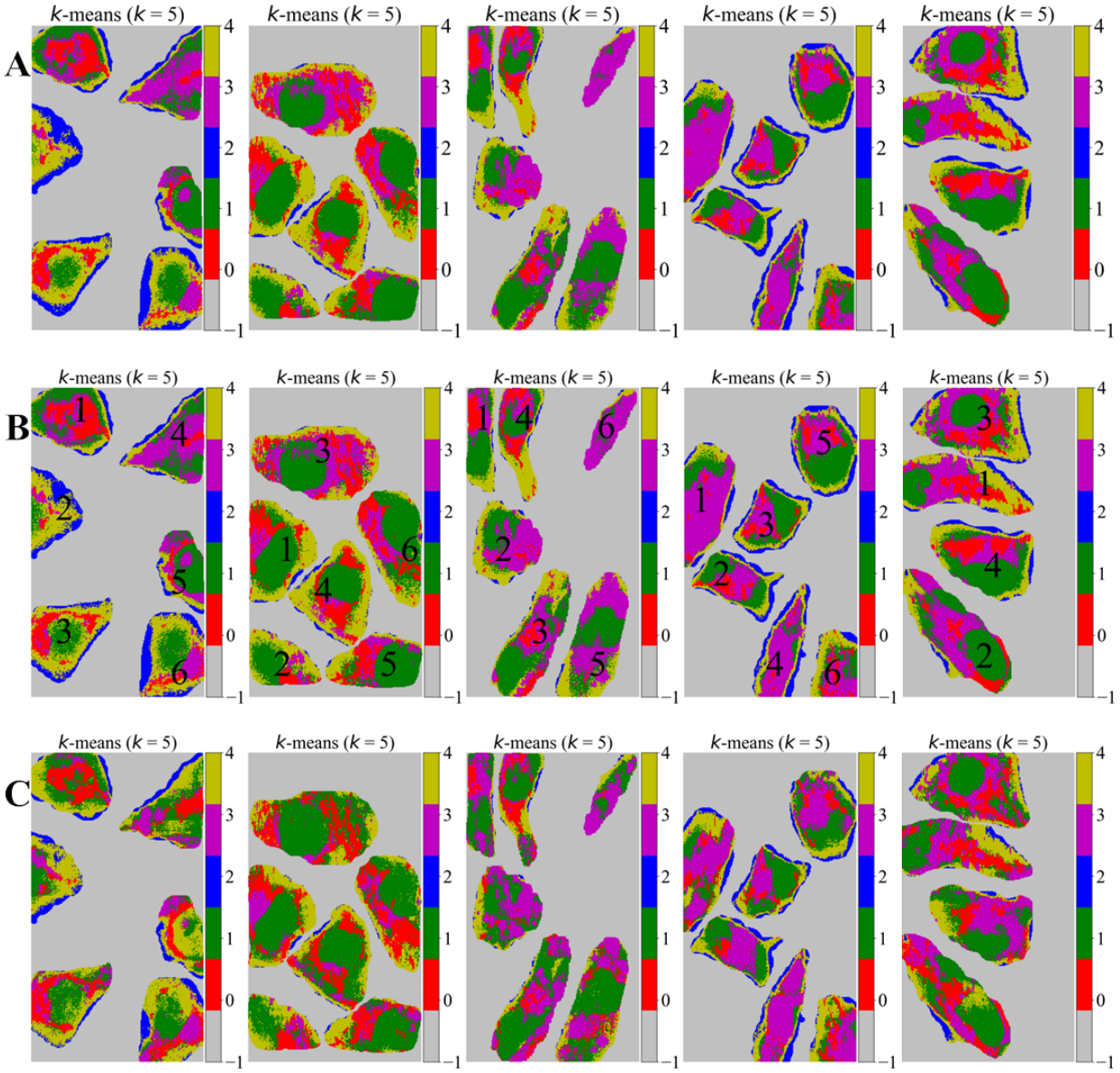

Figure S5: The k-means clustering maps with $k = 5$ for individual Raman spectra of 5 FTC Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

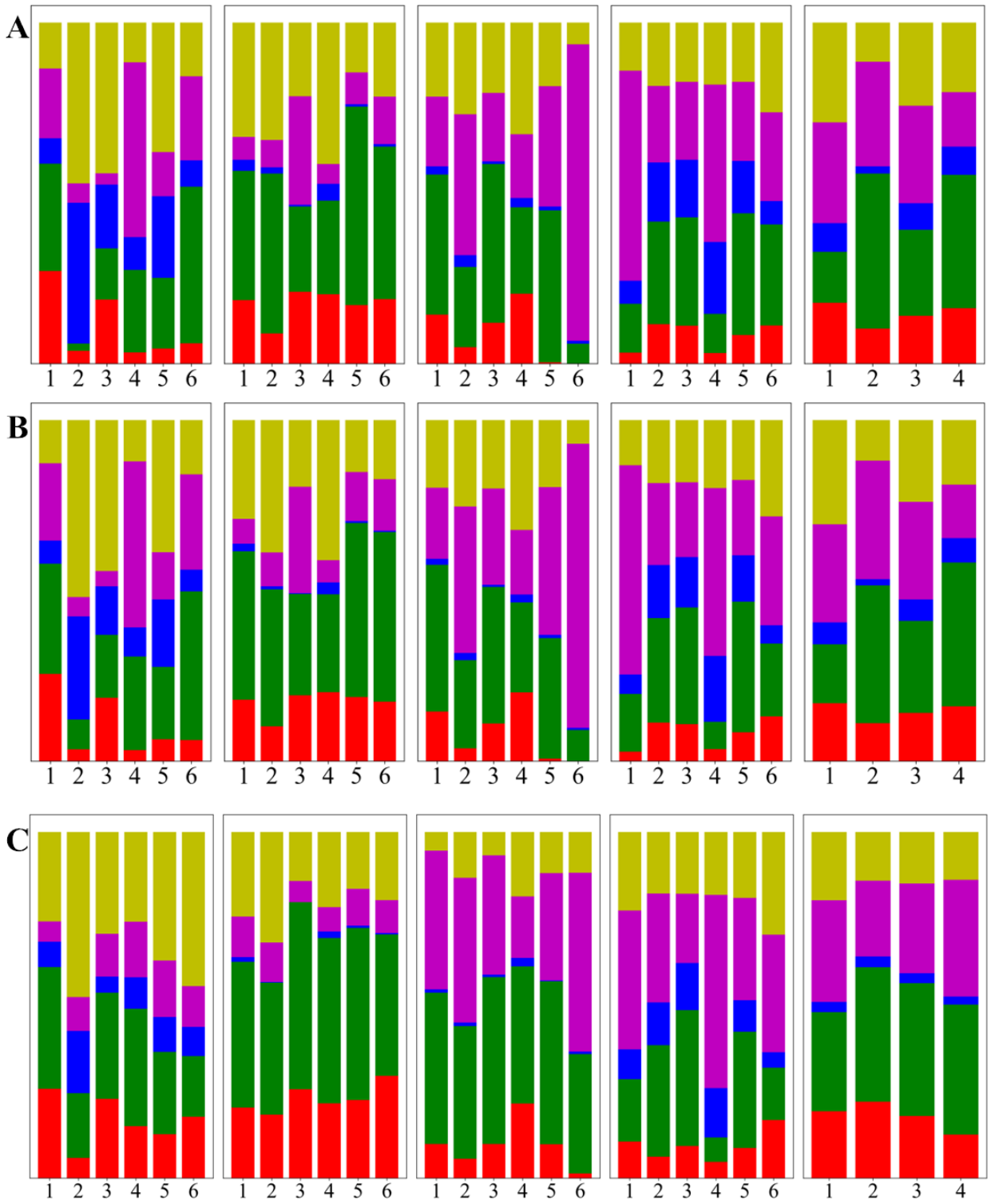

Figure S6: The relative populations of the clusters within each single cell for 5 FTC Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

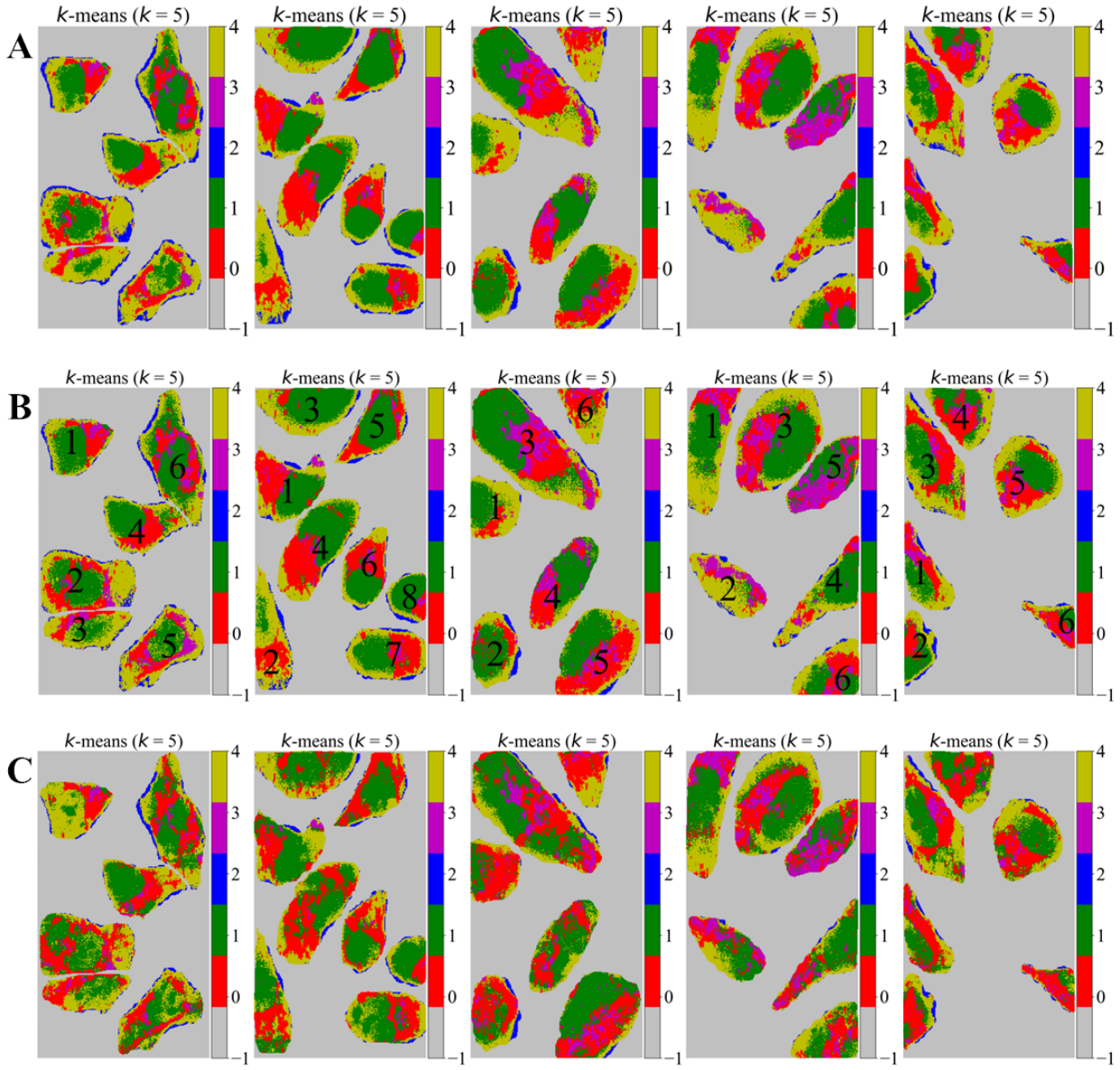

Figure S7: The k-means clustering maps with $k = 5$ for individual Raman spectra of 5 Nthy Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

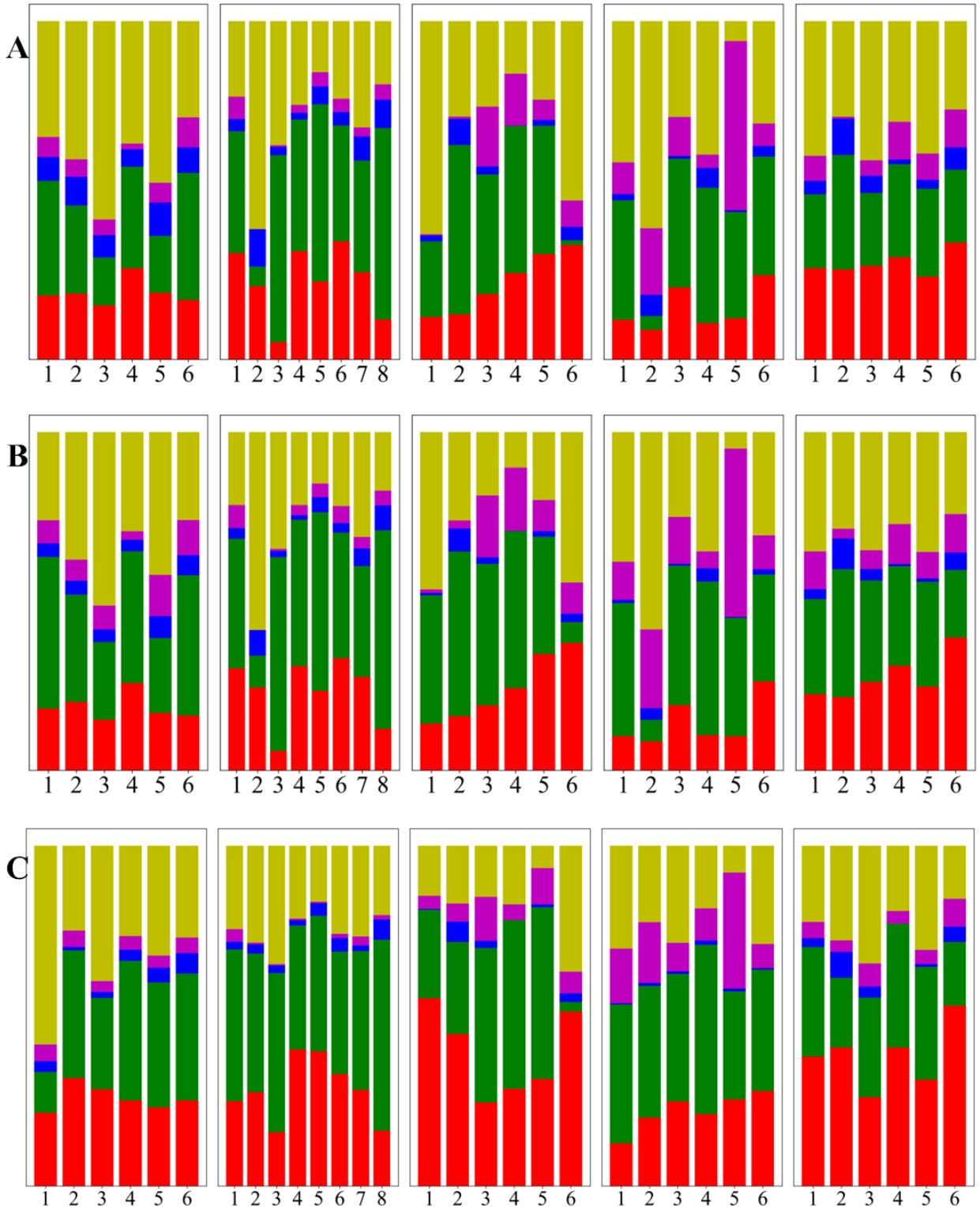

Figure S8: The relative populations of the clusters within each single cell for 5 Nthy Raman images (A) without wavenumber calibration (top row), (B) the position-dependent wavenumber calibration (middle row), (C) the detrending scheme based on random forest regression (bottom row).

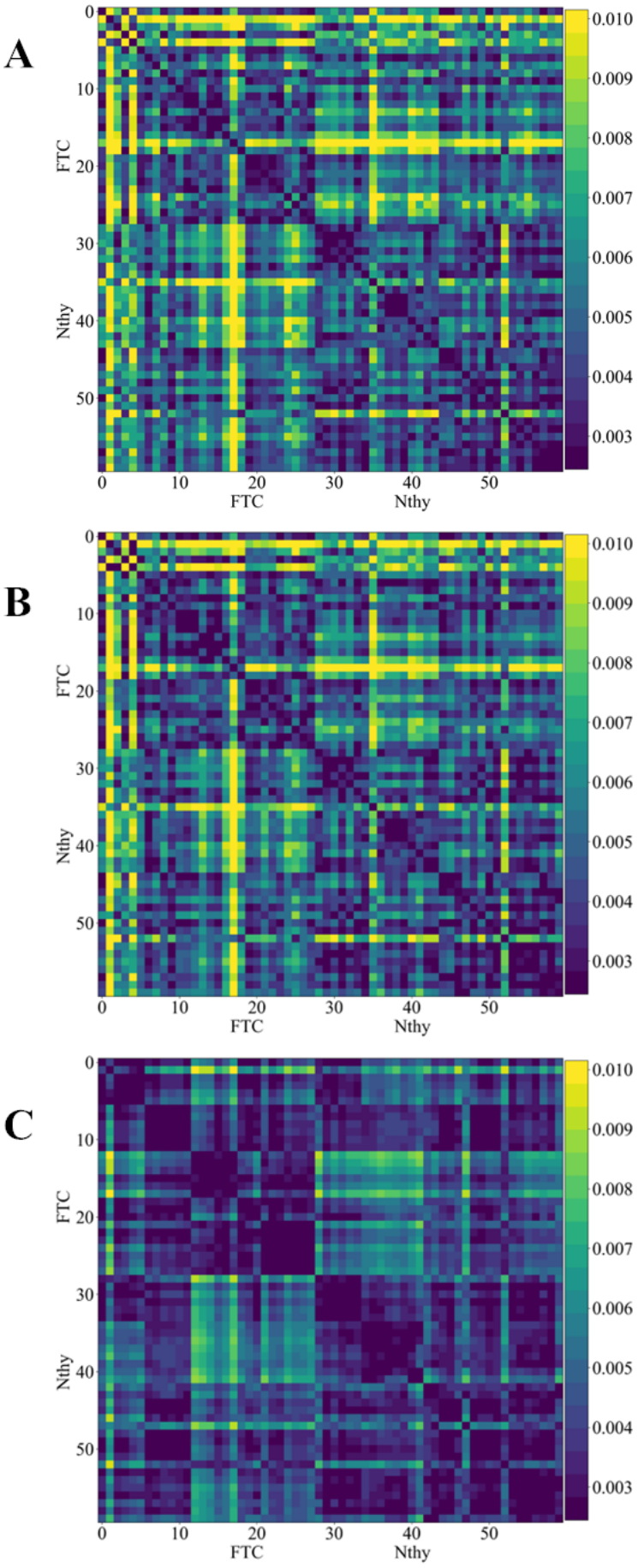

Figure S9: The distance matrix between averaged single cell spectra of sixty cells (28 cells of FTC-133 and 32 cells of Nthy-ori 3-1): (A) without wavenumber calibration. (B) the position-dependent wavenumber calibration. (C) the detrending scheme.

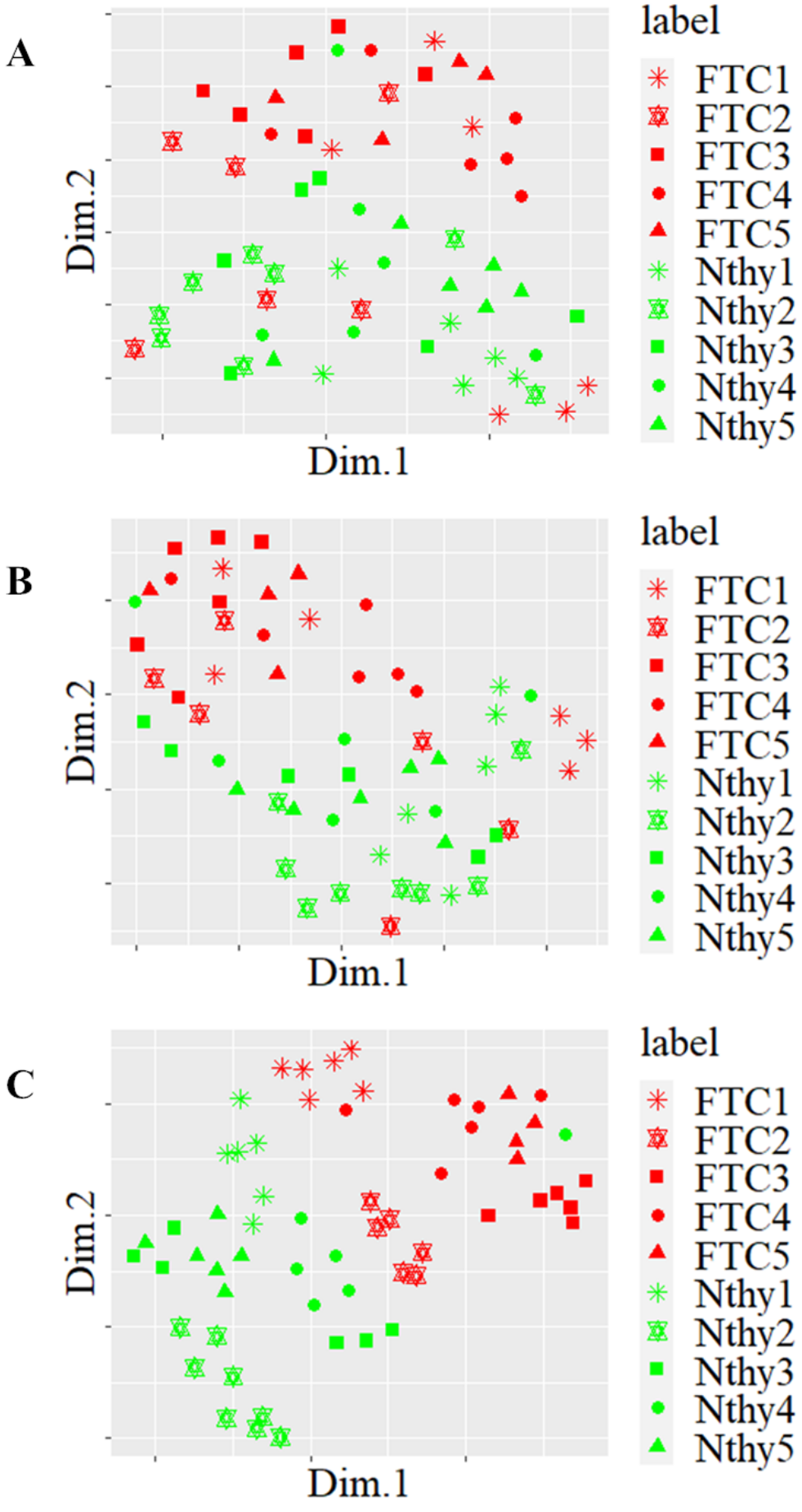

Figure S10: An UMAP projection of averaged single cell spectra of sixty cells. (A) without wavenumber calibration. (B) the position-dependent wavenumber calibration. (C) the detrending scheme.

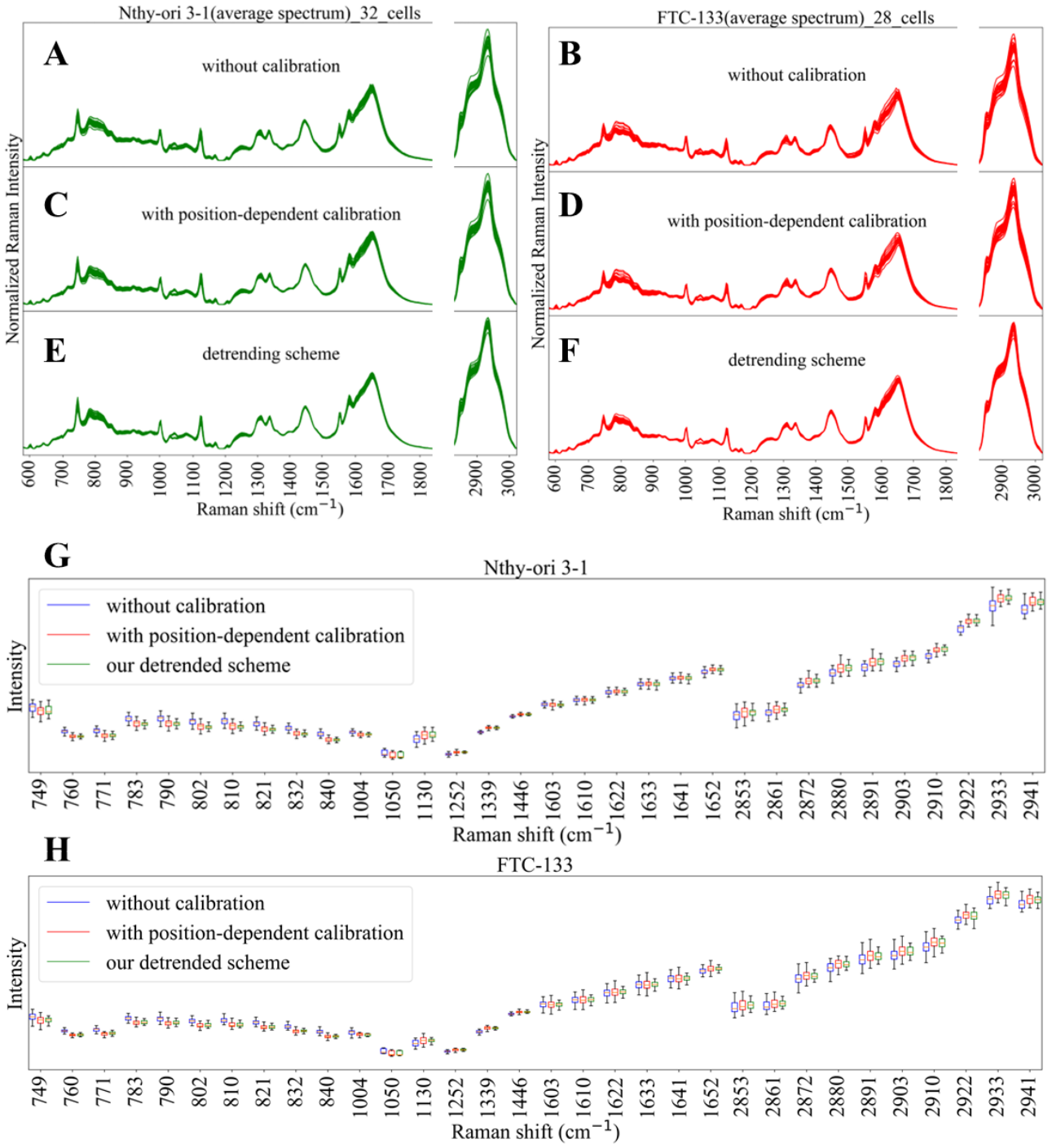

Figure S11: Average spectra of 32 Nthy-ori 3-1 and 28 FTC-133 cells. (A-B) standard preprocessing without wavenumber calibration, (C-D) standard preprocessing with position-dependent wavenumber calibration, (E-F) the detrending scheme after the implementation of position-dependent wavenumber calibration. (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.

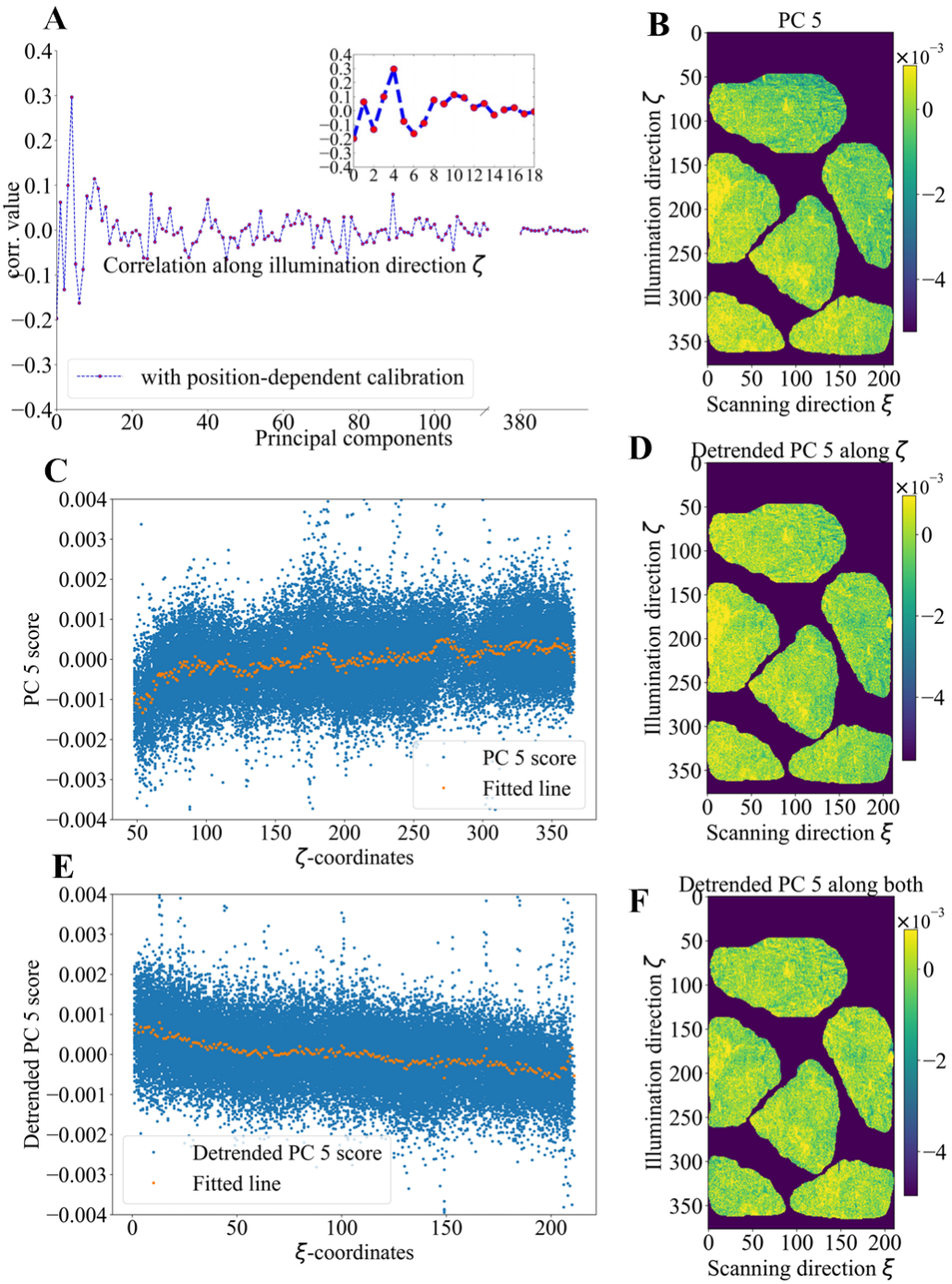

Figure S12: (A) Correlation between illumination axis coordinates and PCs (B) PC 5 scores value distribution in a space domain after standard preprocessing with position-dependent calibration (C) Scatter plot of PC 5 score and ζ-coordinates with RF regression line (D) detrended PC 5 scores value distribution 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.

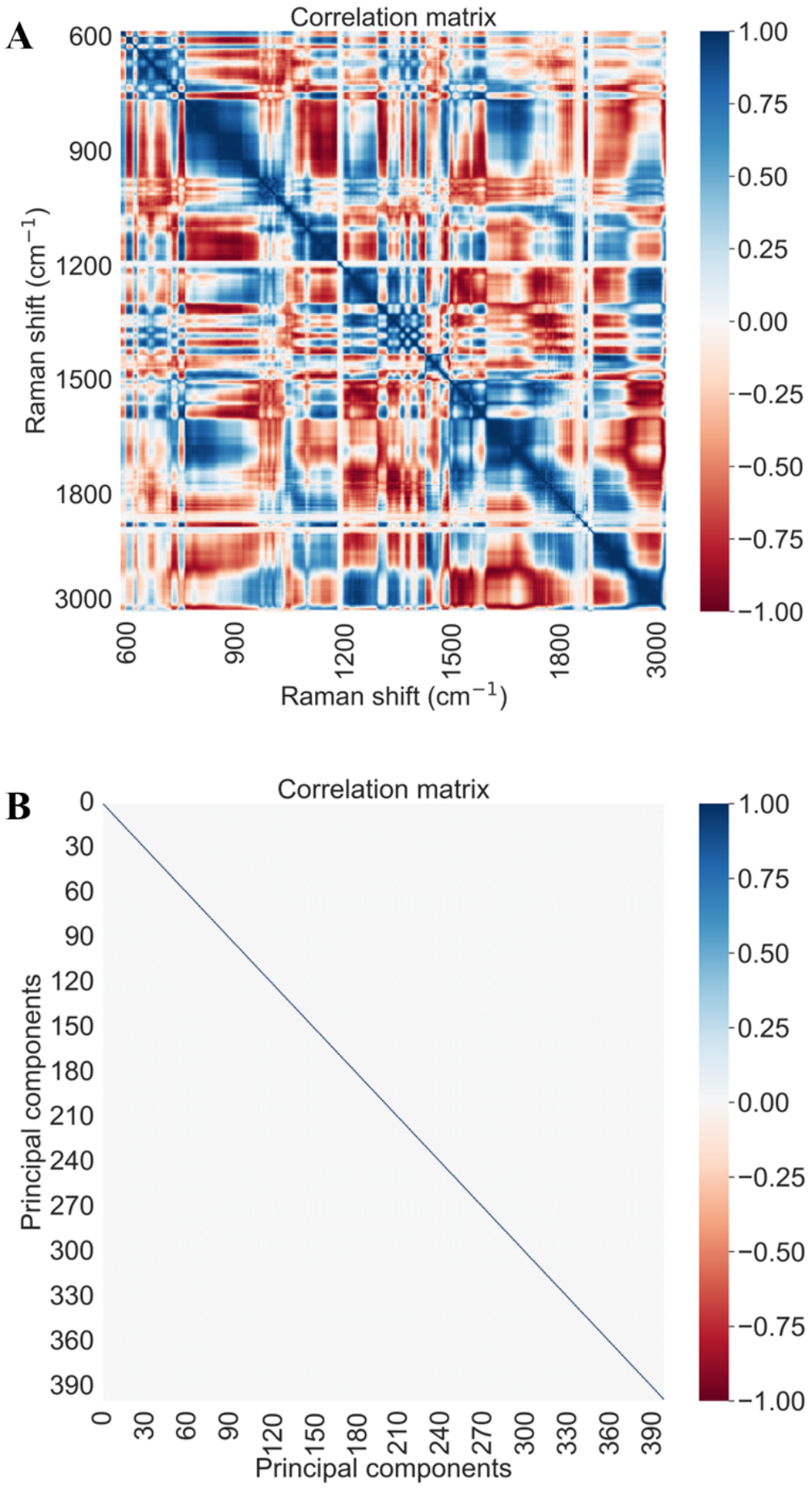

Figure S13: Correlation matrix with position-dependent wavenumber calibration (A) between each Raman shift (B) between each principal component.

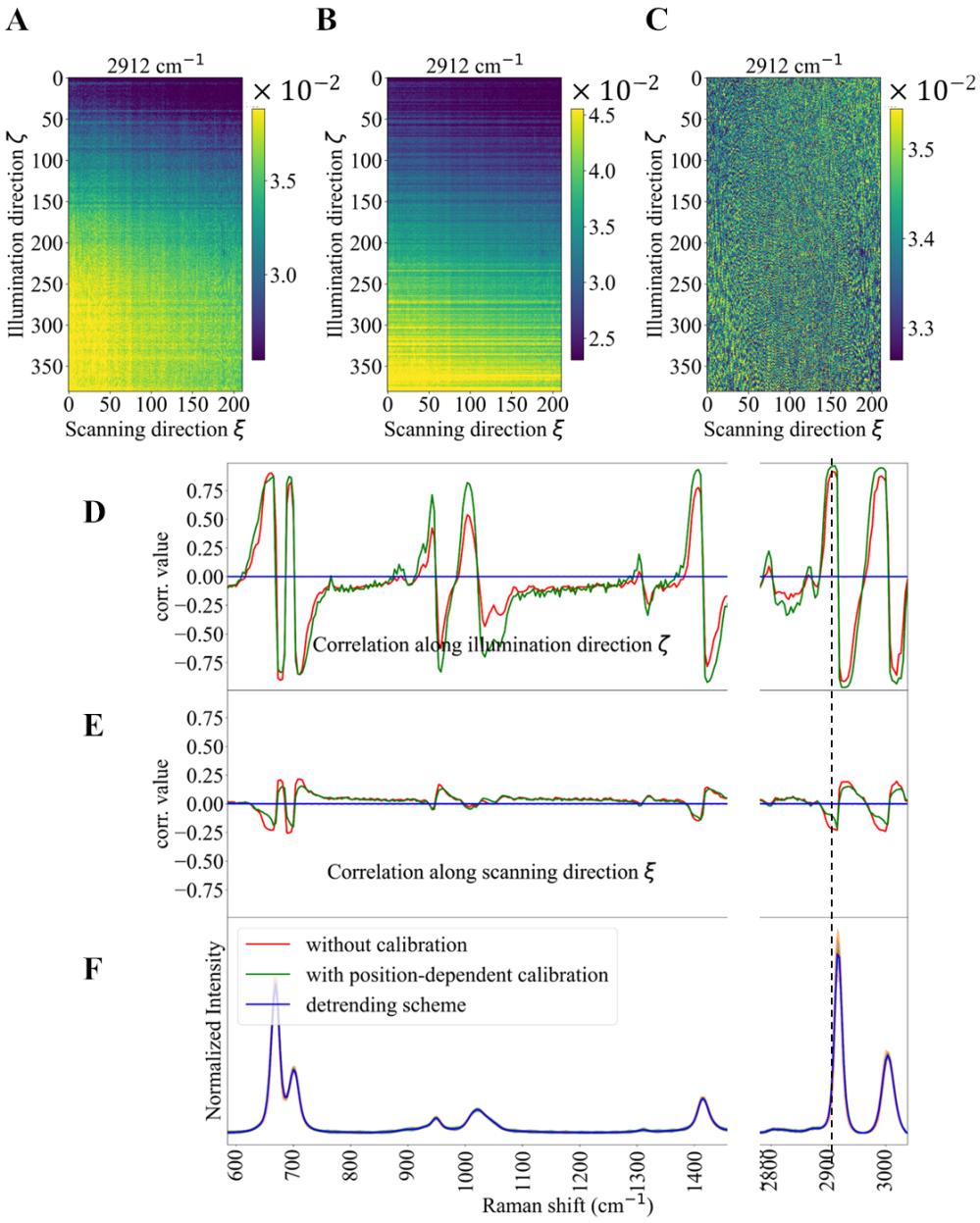

Figure S14: (A)-(C) The Raman intensity distribution at 2912 cm^{-1} (dashed vertical line) in the space domain of DMSO: (A) after standard preprocessing without (position-dependent) wavenumber calibration, (B) after standard preprocessing with position-dependent wavenumber calibration, (C) after the 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.

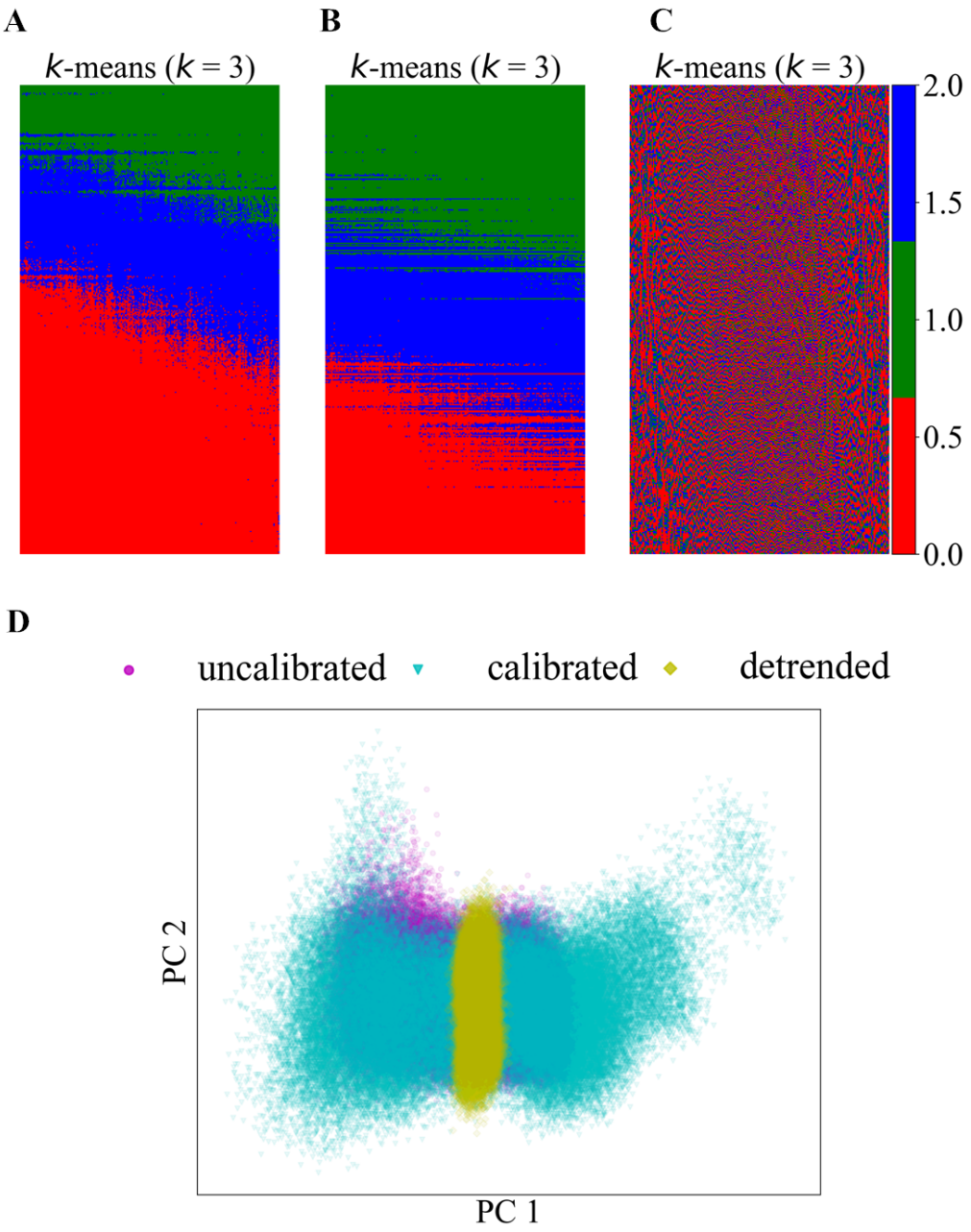

Figure S15: (A)-(C) The k-means clustering maps with $k = 3$ for individual Raman spectra in the Raman image for DMSO: (A) standard preprocessing without position-dependent wavenumber calibration (B) the position-dependent wavenumber calibration (C) the detrending scheme. (D) PCA projection of all spectra based on three preprocessing schemes.

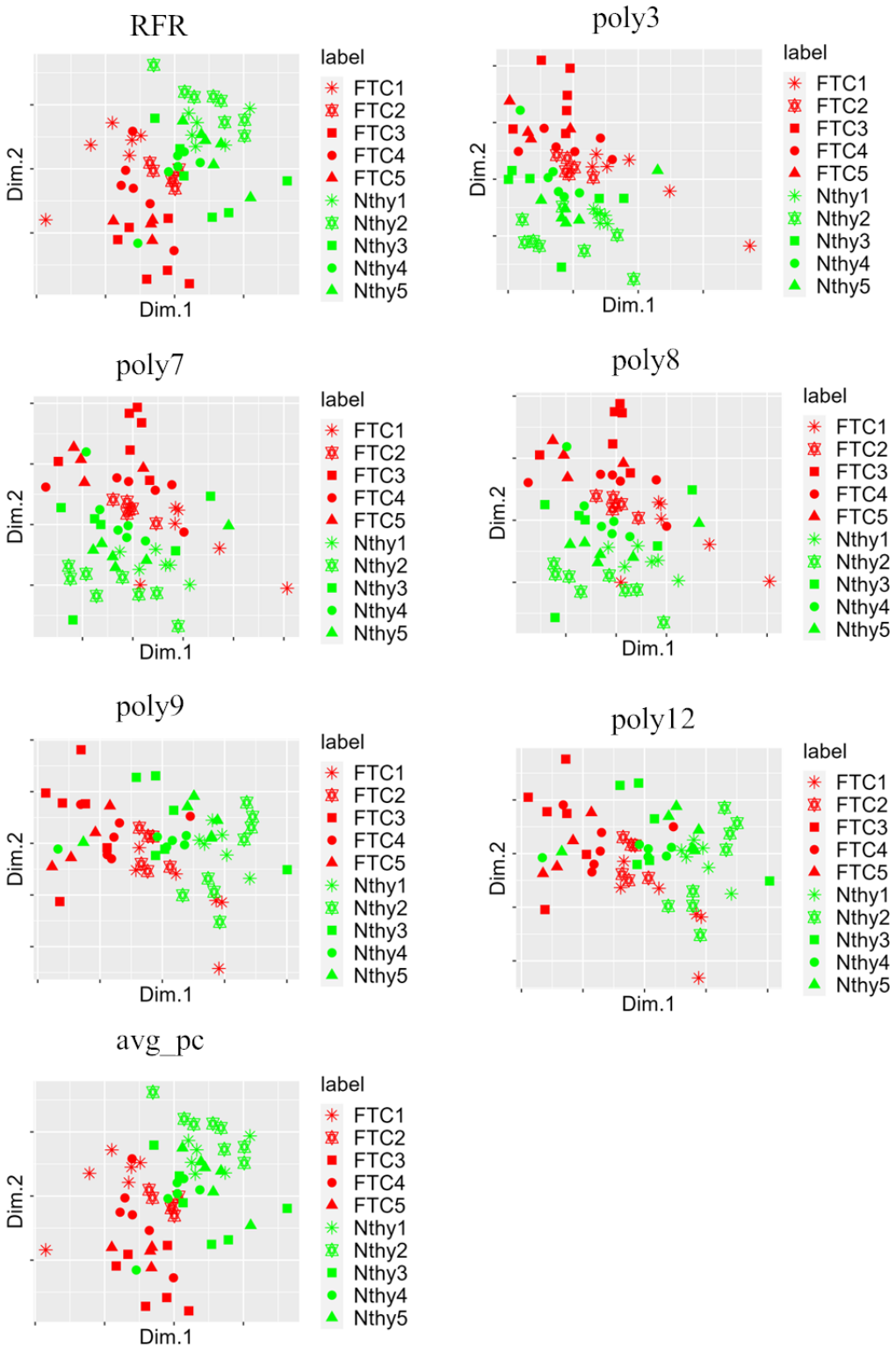

Figure S16: The multi-dimensional scaling (MDS) projection of the ten Raman images including sixty single cells in total for seven detrending methods.

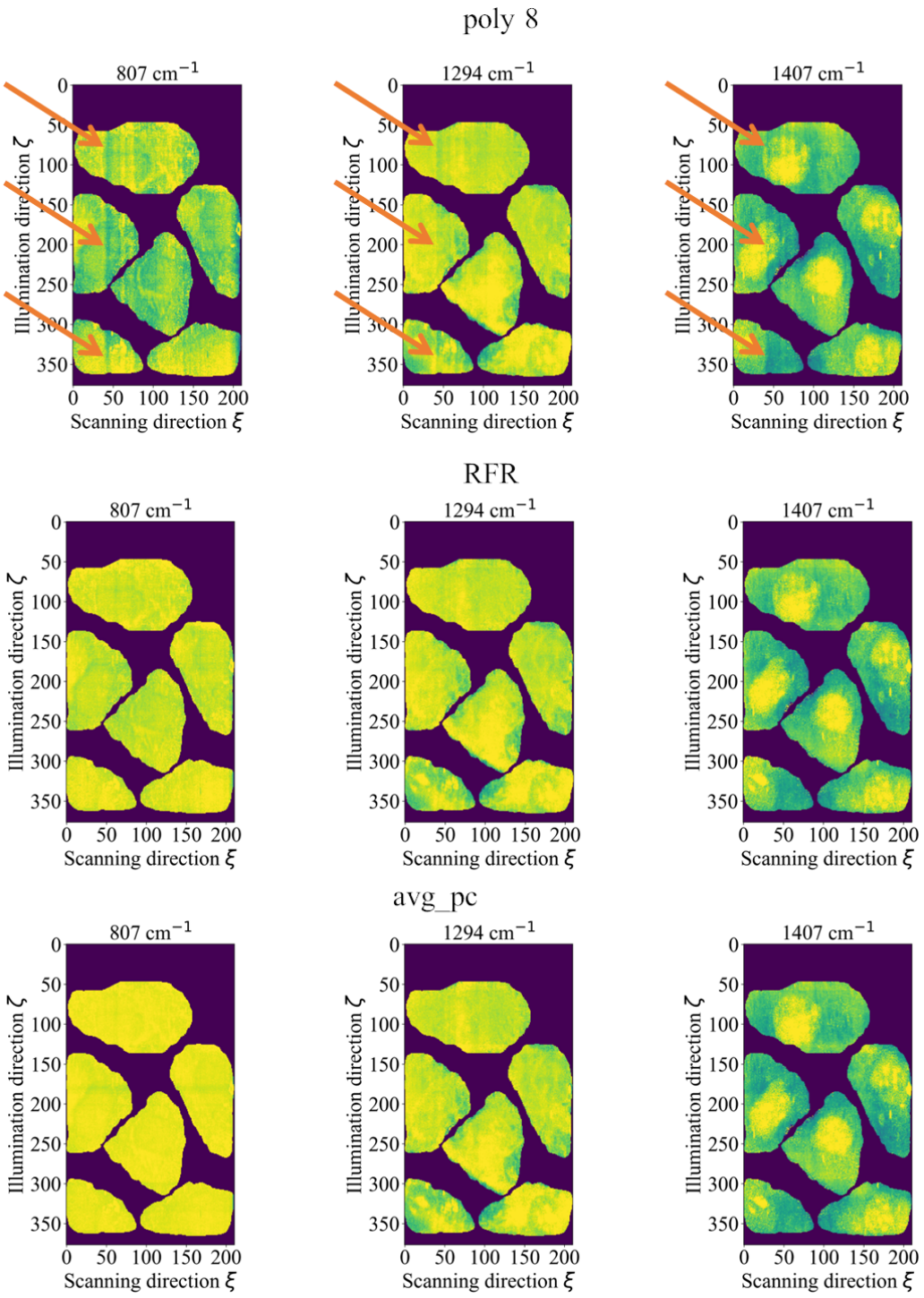

Figure S17: The Raman intensity distribution at three different peaks 807 cm^{-1}, 1294 cm^{-1}, 1407 cm^{-1} in the space domain for the Raman image of FTC-133(#2): polynomial fitting of order 8 (top row), Random forest regression (middle row), average PC (bottom row). Orange arrows indicate that some stripes (scars) persist to exist in the reconstructed Raman images by the polynomial regression scheme.

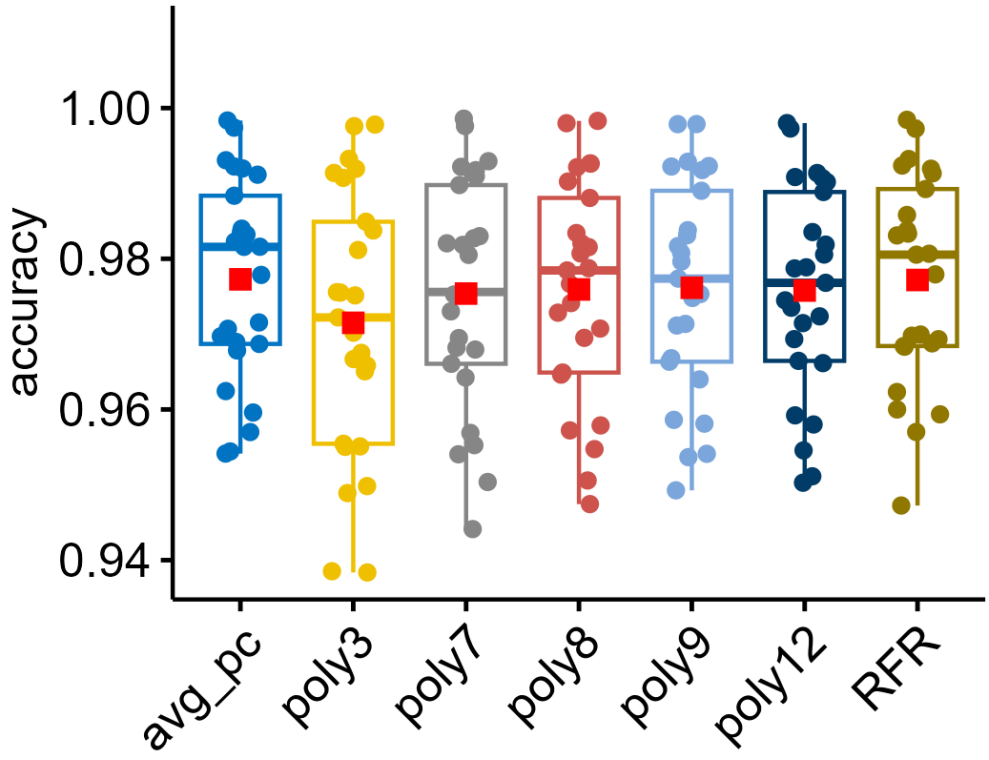

Figure S18: The box-and-whisker plot of test accuracy in the prediction of FTC-133/Nthy-ori 3-1 for seven detrending methods of 25-fold cross validation based on pixelwise spectra.

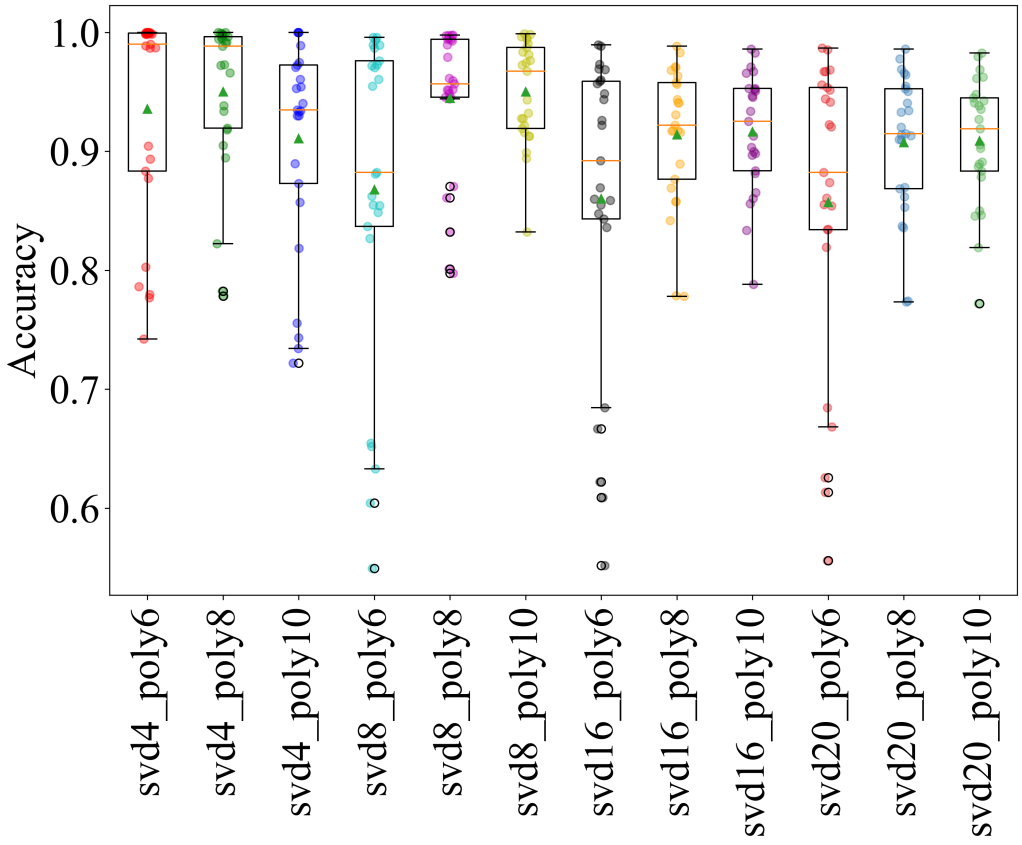

Figure S19: The box-and-whisker plot of test accuracy in the prediction of FTC-133/Nthy-ori 3-1 for the standard preprocessing without wavenumber calibration of 25-fold cross validation based on pixelwise spectra: different pairs of singular value decomposition components for denoising and polynomial fitting orders for baseline corrections.

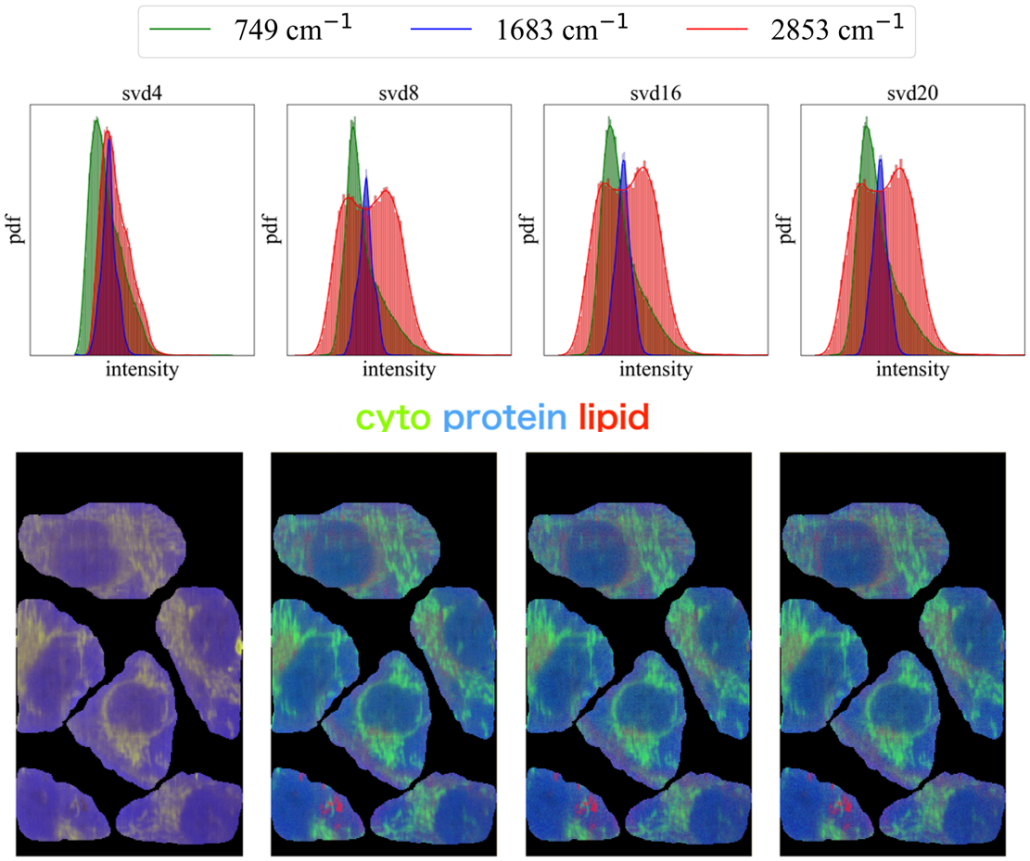

Figure S20: The Raman intensity distribution at known cytochrome peak 749 cm^{-1}, protein peak 1683 cm^{-1} and lipid peak 2853 cm^{-1} for the Raman image of FTC-133(#2) for different singular value decomposition components.

[4] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo and M. Chica-Rivas, *Ore Geology Reviews*, 2015, **71**, 804–818.

[5] H. Ishwaran, *Machine Learning*, 2015, **99**, 75–118.

[6] J. N. Taylor, K. Mochizuki, K. Hashimoto, Y. Kumamoto, Y. Harada, K. Fujita and T. Komatsuzaki, *The Journal of Physical Chemistry B*, 2019, **123**, 4358–4372.

[7] R. Nakayama, K. Horiuchi, M. Susa, S. Hosaka, Y. Hayashi, K. Kameyama, Y. Suzuki, H. Yabe, Y. Toyama and H. Morioka, *Thyroid*, 2012, **22**, 200–204.

[8] Z. Movasaghi, S. Rehman and I. U. Rehman, *Applied Spectroscopy Reviews*, 2007, **42**, 493–541.

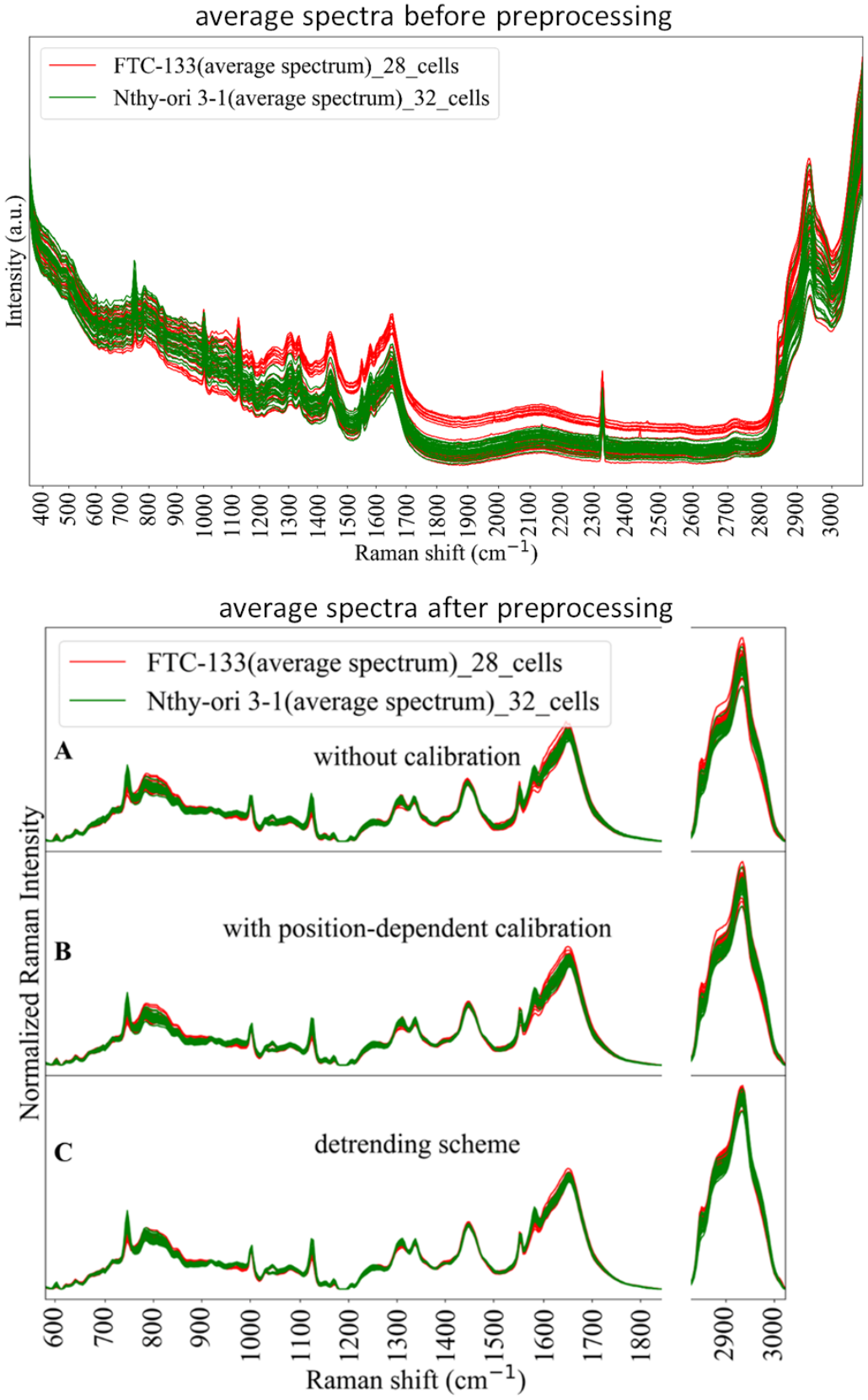

Figure S21: Average spectra of 32 Nthy-ori 3-1 and 28 FTC-133 cells: before preprocessing (top) (A-C) standard preprocessing without wavenumber calibration, standard preprocessing with position-dependent wavenumber calibration,the detrending scheme.

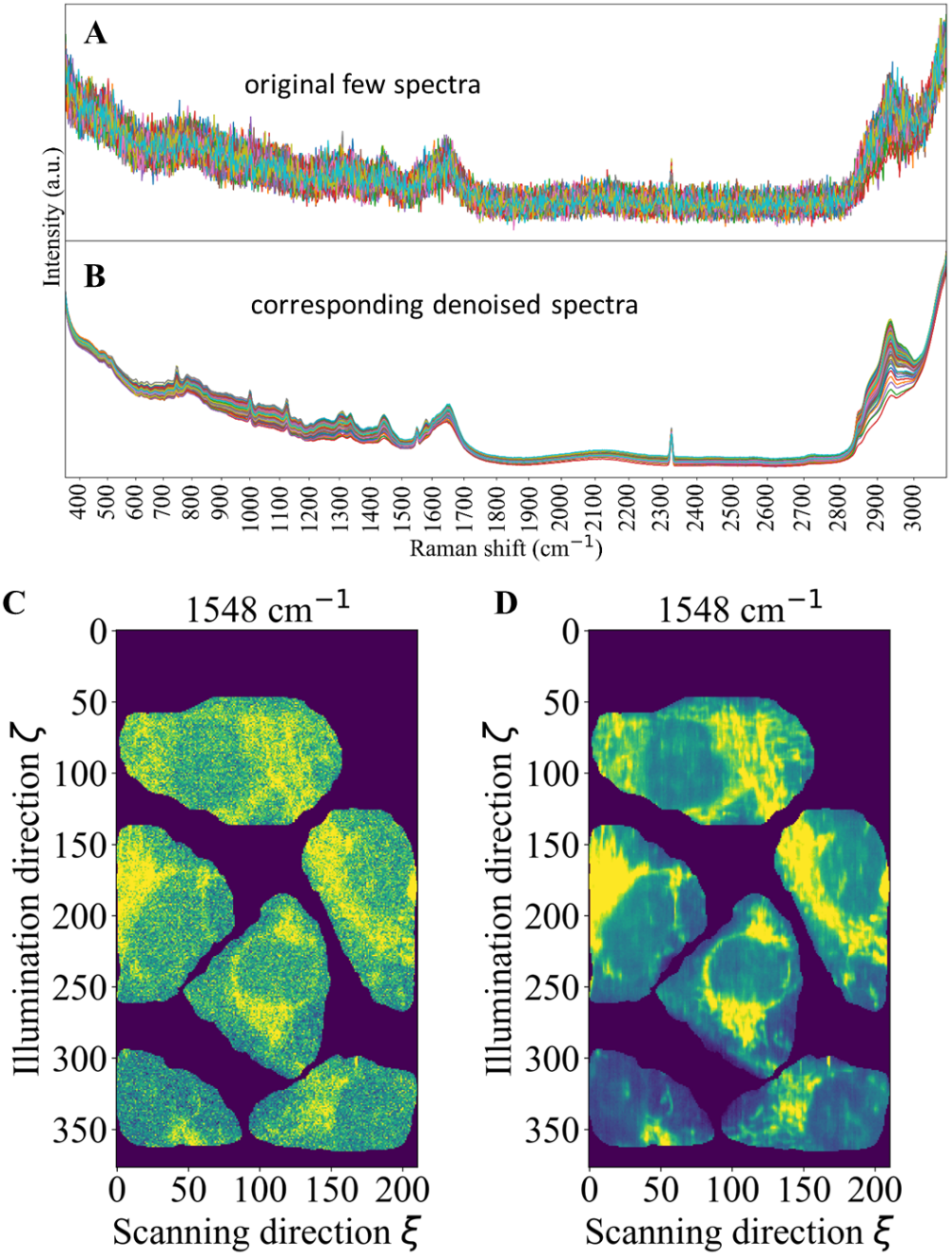

Figure S22: For the Raman image of FTC-133(#2) (A) few individual pixel level raw Raman spectra (B) corresponding denoised spectra, (C-D) intensity distribution at 1548 cm^{-1}: (C) raw Raman image, (D) denoised Raman image.

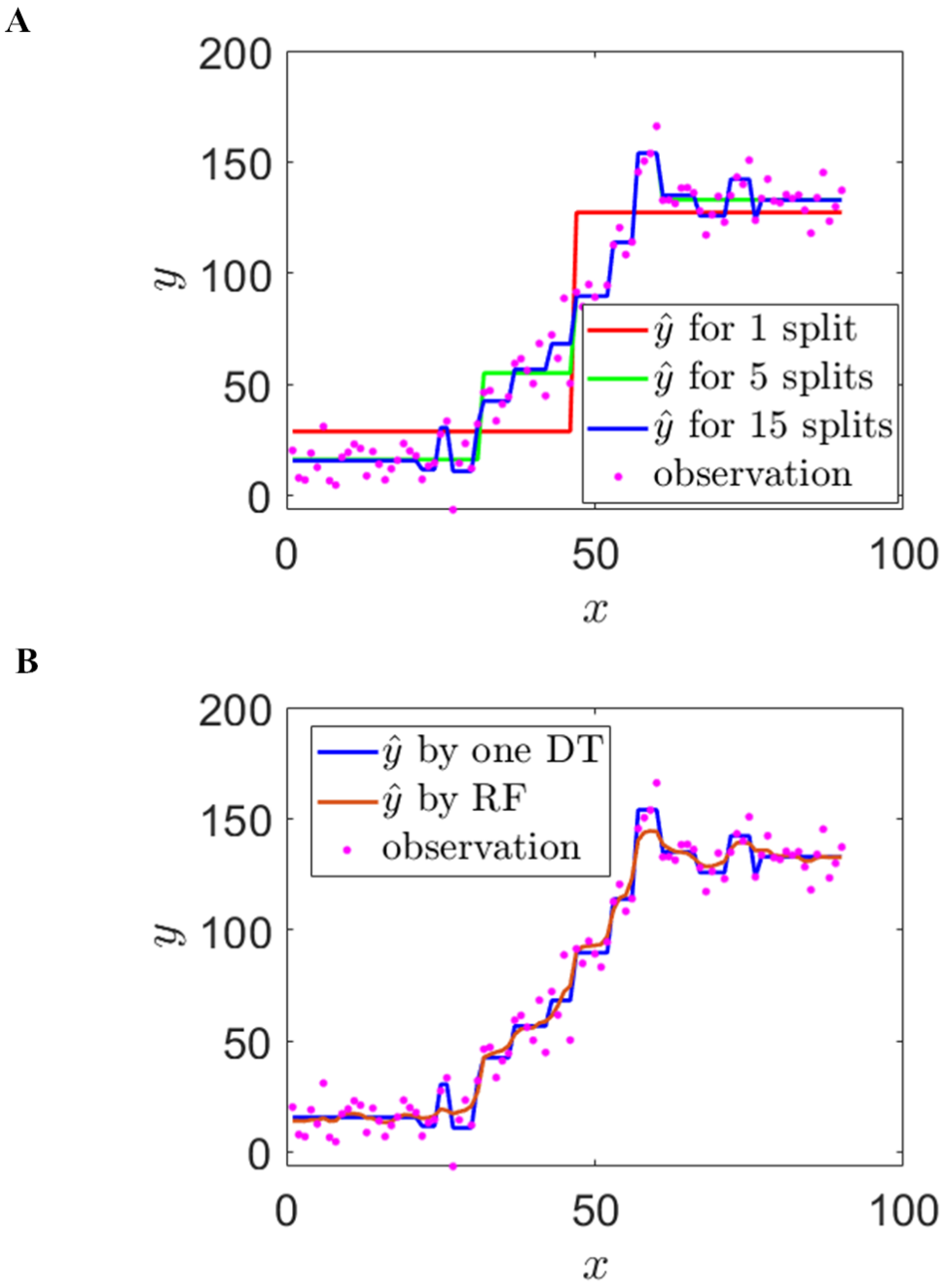

Figure S23: (A) A scatter plot of "observed" data with the approximate regressions by a DT in terms of 3 different splits (1, 5, and 15). (B) A scatter plot with the approximate regressions by a single DT with 15 splits and by a set of 100 DTs.

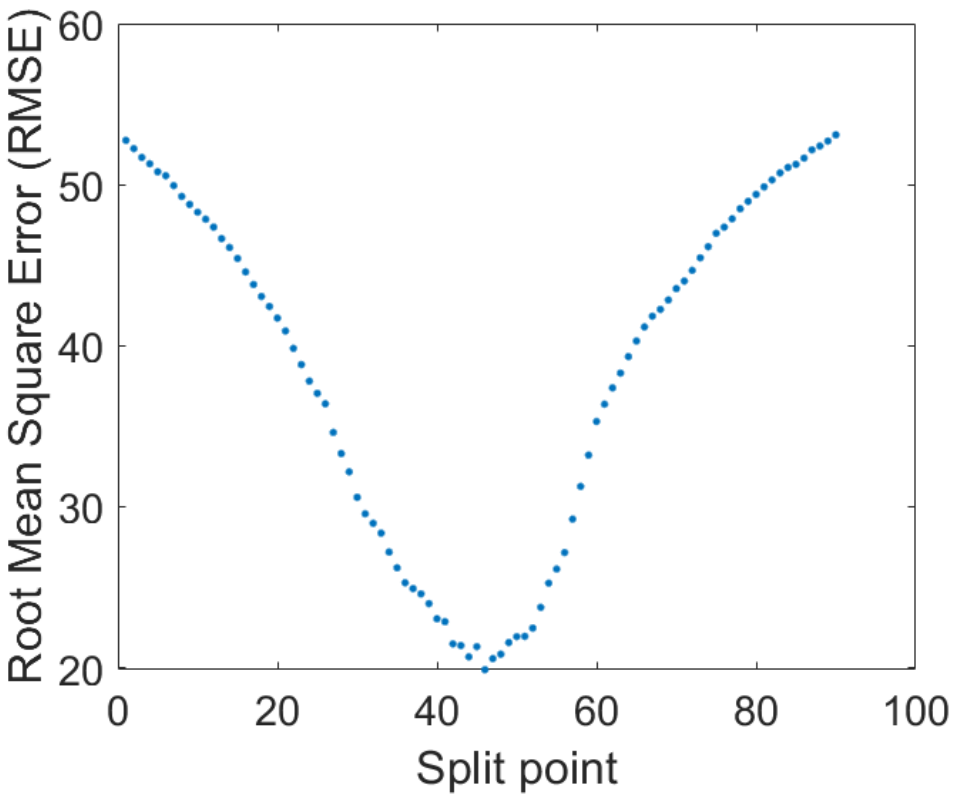

Figure S24: The root mean squared error (RMSE) as a function of the location at which the point to split the given data set is determined for the data set used in Fig. S23.