Tutorial for the Rametrix™ LITE Toolbox v1.0

This tutorial is also contained in the Supplementary Appendix of our original Rametrix™ publication: Fisher et al. (2018) The Rametrix™ LITE Toolbox v1.0 for MATLAB®. Journal of Raman Spectroscopy. 49(5):885-896.

The following section contains descriptions and example figures of each of the tabs of the Rametrix™ LITE Toolbox taken during the bacterial growth experimental analysis. The bacterial growth spectra are provided online (see below) so new users may acquaint themselves with the Rametrix™ LITE Toolbox functionalities.

Rametrix™ LITE Toolbox Availability and Requirements

The Rametrix™ LITE Toolbox is freely-available to academic users through GitHub (https://github.com/SengerLab/RametrixLITEToolbox). The bacterial growth spectra can also be obtained from this site. The Rametrix™ LITE Toolbox was developed in MATLAB® R2016a and requires the Bioinformatics Toolbox™ and Statistics and Machine Learning Toolbox™. The Rametrix™ LITE Toolbox can be run on PC and Macintosh.

Install and Launch the Rametrix™ LITE Toolbox in MATLAB®

The Rametrix™ LITE Toolbox can be installed in MATLAB® by first downloading the 'Rametrix(TM) LITE.mltbx' file from the GitHub folder given above. Move the file to the desired location in your file system, and double-click the file to start MATLAB® and the install dialog box. From here, click 'install'. To start the Rametrix™ LITE Toolbox in MATLAB®, type 'Rametrix' in the command window.

Start Module

The Start module, shown in Figure S1, contains four buttons for loading Raman spectra in different file formats and three buttons for saving program output in various formats. All load buttons, except for the Load .RDA button, will request a folder location containing all spectra for analysis in the indicated file format. The spectra will be internally labeled by factors given in the file names, with individual factors separated by underscores (i.e. 'ecoli_2hours_growth.spc'). All spectra must have the same number of factors.

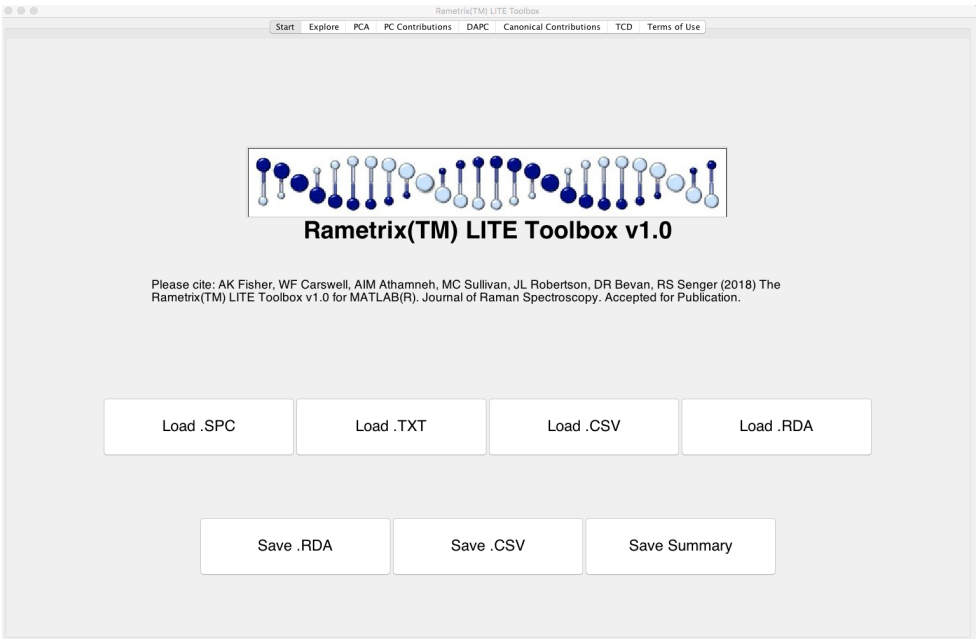

Figure S1. The Rametrix™ LITE Toolbox Start module.

The Load .SPC button will read .SPC files, a standard file format for many spectroscopic instruments. The Load .TXT button can read plain text files containing the wavenumbers and the intensities in two columns with only whitespace separating the values. The .CSV file format should be used for comma separated values with each row containing a wavenumber followed by its associated intensity, separated by a comma. The .RDA file format is unique to the RDA Toolbox and when loaded will restore a saved session in the Toolbox.

The Save .RDA button will preserve the settings of an Rametrix™ LITE Toolbox work session to be recalled later. The Save .CSV button will save multiple comma separated value files prefixed by what preprocessing manipulation was performed on the spectrum contained in each file. When in a PC operating system, the Save Excel Summary button will create a Microsoft Excel® formatted file with separate work sheets for the information given by each Rametrix™ LITE Toolbox module (located in tabs). In a Mac operating system, the Save Excel Summary button will instead request a folder location in which to save multiple .csv files that will contain the information from each Rametrix™ LITE Toolbox module. Writing the summary will take several moments on most systems.

Explore Module

Once the spectra have been loaded, the table on the left of the Explore module tab will populate as shown in Figure S2, with each column in the table containing the sequential factors that were separated by underscores in each spectrum file name. The first column in the table contains checkboxes that, when enabled, will cause the associated spectrum to be disregarded in the PCA and DAPC module calculations. Selecting a spectrum or multiple spectra in the table will display their data in the figure window in the center of the screen, modified by the option selected in the popup menu along the bottom of the screen.

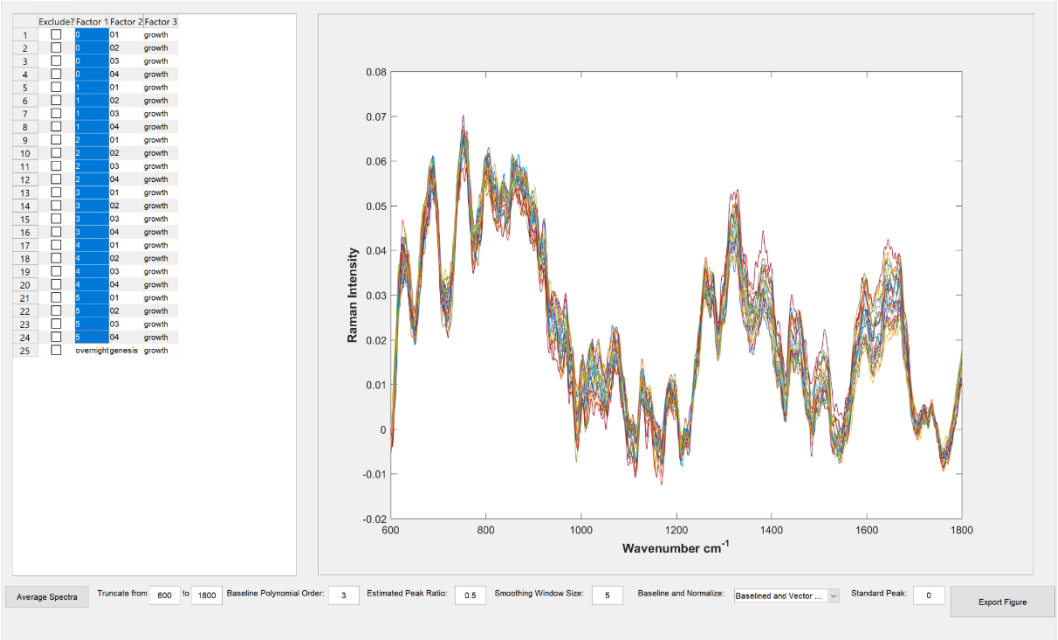

Figure S2. The Explore module.

Directly beneath the spectra table is the Average Spectra button. Activating it will initialize the calculation to average the raw intensities of each set of spectra of *n* factors when the first *n-1* factors match. It will then delete the last column in the table, which is recommended to contain replicate number of a sample scan. If any spectra are enabled for exclusion, the Average Spectra functionality will delete them before performing its calculation.

Along the bottom of the tab are text fields in which to enter the desired wavenumber range for analysis followed by text fields describing the desired inputs to the Goldindec algorithm, the baseline polynomial order, and estimated peak ratio. If baselining calculations are unable to coalesce within 30 iterations, the algorithm is forced to continue on to the next spectrum. Baseline shape for every spectrum should be reviewed before proceeding with analysis. Window size for smooth, a MATLAB® function to smooth spectra, can be specified via the fifth text field.

A popup menu enables selection of the desired display option for spectra selected in the table, and will apply the selected operation to all loaded spectra. The first option, 'None', will only apply truncation to the loaded spectra. The second option, 'Show Baseline', is primarily used to preview the effects of the Goldindec algorithm on the truncated spectra. The third option will display the selected baselined spectra, and the fourth and fifth options will apply the selected normalization method to the baselined spectra. Vector normalization is defined in Eq. 1 in the main article.

Wavenumber normalization divides each Raman intensity in the truncated range of a spectrum by the baselined intensity in the same spectrum at the standard peak specified by the user in the remaining text field. The selected standard peak must be inside the truncation range.

The Export Figure button will export the current RDA Explore tab graph into a MATLAB® figure window for ease of customization, analysis, and saving the figure in different file formats.

PCA Module

The principal component analysis (PCA) module, shown in Figure S3, displays the same table of spectra and factors as the Explore tab and will dynamically update as spectra are selected for exclusion or inclusion in any following calculations. The Run PCA button will perform principal component analysis on all spectra, except on those that have been designated for exclusion, using the 'pca' MATLAB® function. An input (*n* x *p*) data matrix of Raman intensities from each spectrum *n* by wavenumber *p* will be orthogonally transformed by the 'pca' function.

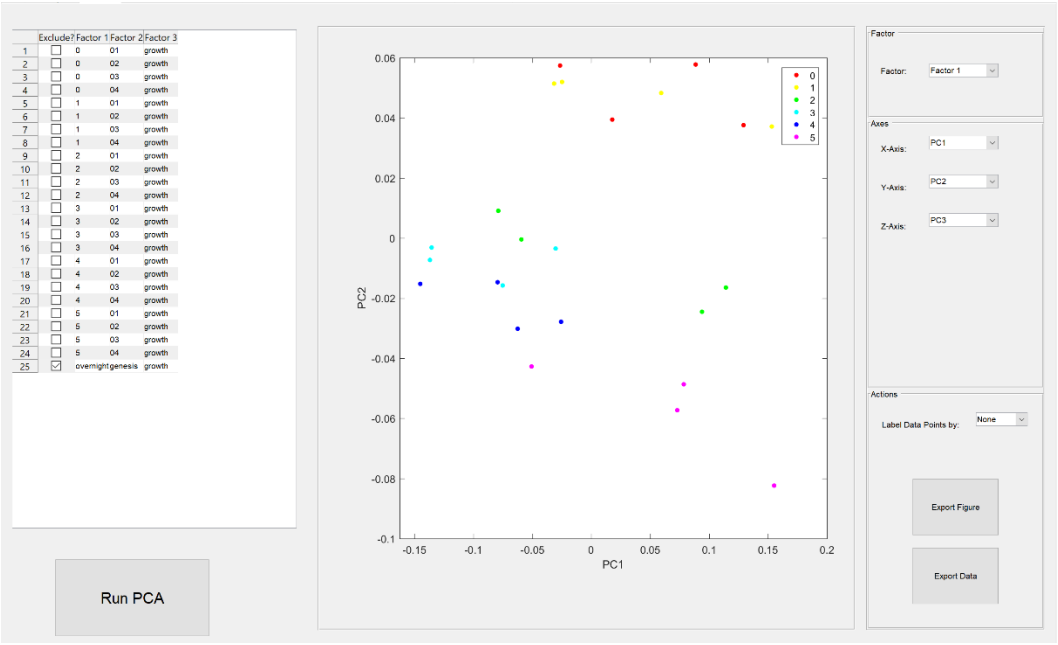

Figure S3. The PCA module.

The input matrix will be comprised of the intensities adjusted by user specifications in the Explore tab popup menu. If the 'None' option is selected, then only the raw spectra in the truncation range will undergo PCA. The 'Show Baseline' and 'Baselined Only' options will provide the 'pca' function with the baselined wavenumber intensities within the truncation range. The 'Baselined and Vector Normalized' and the 'Baselined and Wavenumber Normalized' options will provide the latest normalized spectra within the truncation range to the 'pca' function.

The total number of principal components will be one less than the number of spectra provided or equal to the number of wavenumbers provided in the spectra, whichever value is lower. Projecting the Raman intensities onto the principal components transforms the intensity values into principal component scores and allows each spectrum to be graphed as a single data point along the principal components.

The figure in the center of the PCA module will default to an X-Y-Z coordinate system where each axis is one of the first three principal components calculated by the Run PCA button functionality. On the right of the PCA module, a specific factor position can be selected. This will not affect any calculations done by the Run PCA button, but will determine the initial data point color, legend, and potential labeling in the generated figure. The Run PCA button must be activated again for any change to the data point color or figure legend by the factor selection to

take effect. Each axis can be adjusted to display the PCA Scores for a different principal component using the X, Y, and Z-axis popup menus, and the figure itself can be rotated for a better view. The last popup menu on the right of the tab allows the user to select among three different options for labeling data points: unlabeled, by row in the table on the left of the tab, or by the subcategory of the currently selected factor. The generated figure can also be exported to MATLAB® as before in the Explore module. Finally, the Export Data button will direct the user to save a .CSV file containing a matrix of PCA scores.

PC Contributions Module

After the Run PCA button in the PCA module is activated, the MATLAB® 'pca' function output is used to calculate the fractional contribution of each wavenumber to each principal component. The 'pca' function assigns coefficients to each wavenumber for every principal component, and the fractional contribution of the variability of each wavenumber in the supplied spectra to each principal component can therefore be calculated and displayed. The Rametrix™ LITE Toolbox calculates the fractional contribution of each wavenumber to each principal component according to Eq. 2 in the main article.

This calculation is carried out separately for every wavenumber and principal component (PC). The table on the left of the PC Contributions module, shown in Figure S4, lists every principal component. Enabling the checkbox next to each principal component will graph it in the figure in the center of the display. The user can then evaluate which wavenumbers were of greatest importance to each principal component. Additionally, the % contribution of each principal component to the total system variability is displayed in the second column of the table.

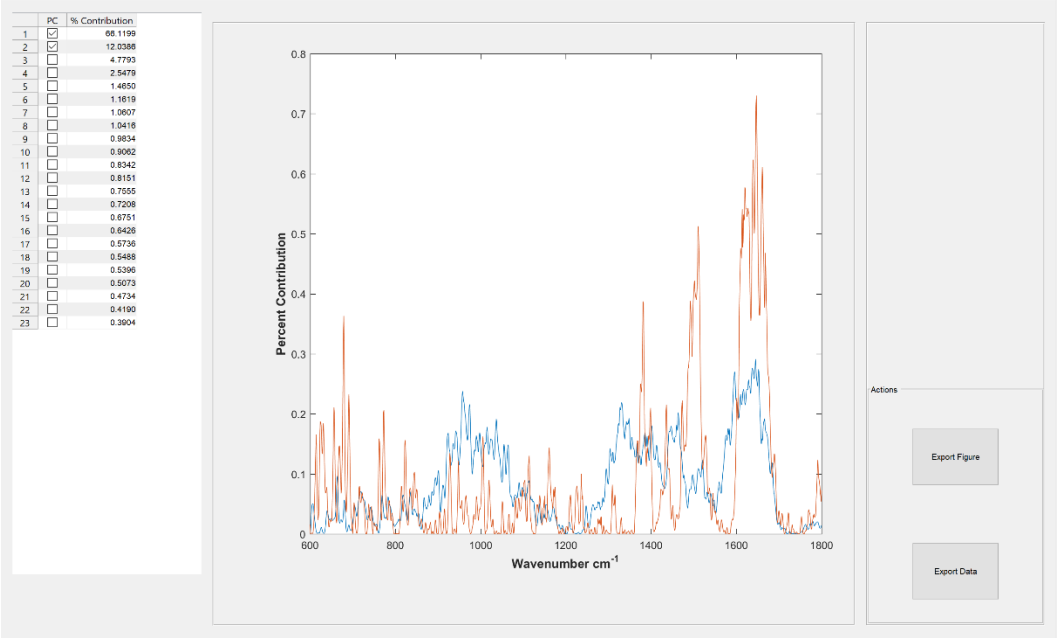

Figure S4. The PC Contributions module.

DAPC Module

The discriminant analysis of principal components (DAPC) module, shown in Figure S5, shares the same table on the left as the Explore and PCA modules and dynamically updates the spectra designated for exclusion. Before activating the Run DAPC button, the user must select the factor with which to define the spectrum groups in the popup menu in the upper right-hand corner of the display. The Rametrix™ LITE Toolbox uses the MATLAB® function 'manova1' to perform discriminant analysis utilizing the user-supplied factor labeling of spectra as *a priori* groups and the % Variability Explained by PCs text field or the Number of PCs Included text field to decide how many PC Scores to pass to the function. Editing the value in either field will update the value in the other.

After adjusting settings and activating the Run DAPC button, the figure in the center of the tab will display the first three canonicals calculated by the 'manova1' function as the X, Y, and Z axes of the graph. The user can then manipulate the display in the same manner as in the PCA module; adjusting axes, labeling data points, rotating the figure, exporting the figure, and exporting the data.

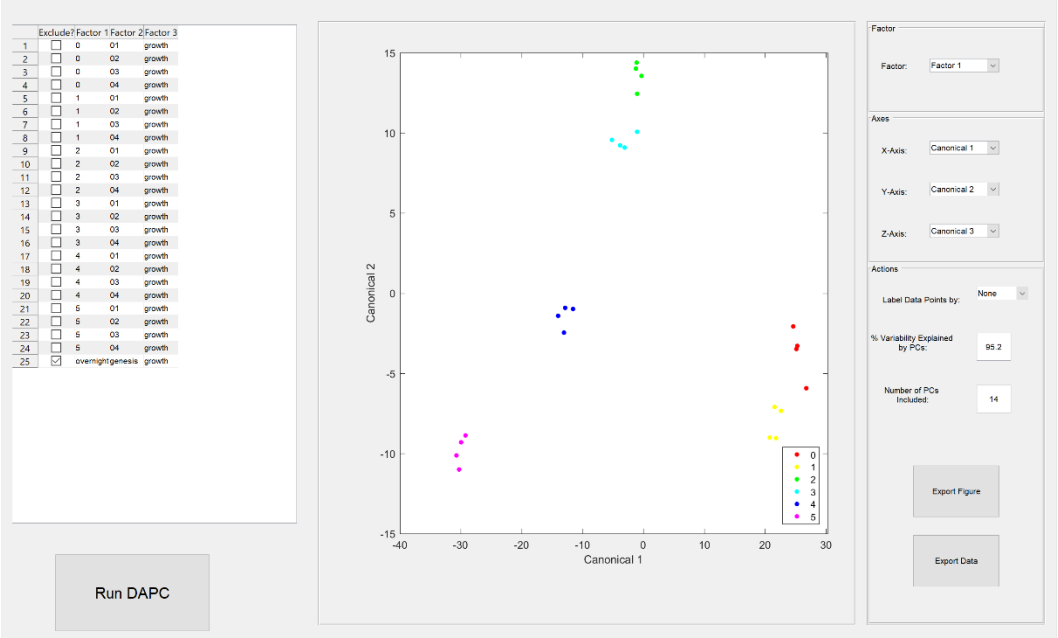

Figure S5. The DAPC module.

Canonical Contributions Module

Like the PC Contributions module, the Canonical Contributions module (Figure S6) calculates the wavenumber contribution to each canonical after the Run DAPC button is activated. The Rametrix™ LITE Toolbox uses Eq. 2 in the main article, this time with *z* representing the loading of each principal component given to the Run DAPC functionality. After the fractional contribution of each principal component is calculated, the Rametrix™ LITE Toolbox multiples the matrix of wavenumber fractional contributions by principal components stored from the Run PCA functionality with the newly made matrix of principal component fractional contributions by

canonicals. The resulting matrix of wavenumber loadings by canonicals is used to populate the data for the table in the Canonical Contributions module, which can be operated in the same manner as the table in the PC Contributions module.

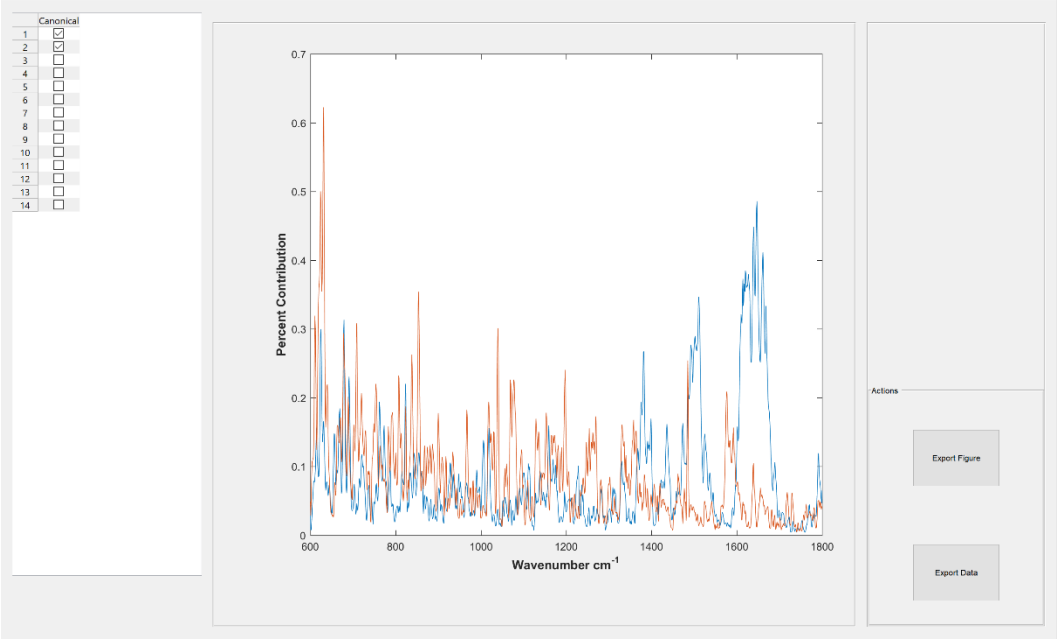

Figure S6. The Canonical Contributions module.

TCD Module

The total canonical distances (TCD) module, shown in Figure S7, provides a quantitative interpretation for the output of the DAPC module. All canonicals, not just those shown in the viewing pane in the DAPC module, are used to calculate the distances of each group from a reference group. The table on the left of the tab lists the groups, provided by the factor selected by the user in the latest DAPC calculation, in its leftmost column. Selecting one of those groups will designate it as the reference group and the total canonical distance calculation will be performed, populating the next two columns in the table. The final column in the table is editable, and the user may input independent variable values for each group into the cells in that column.

Once all cells in the table have been filled, the user can click on the Calculate Correlation button. The 'polyfit' and 'polyval' MATLAB® functions will use the independent variable values entered by the user and the total canonical distances calculated by selecting a reference group to calculate the linear fit between the data. The resulting line will be displayed along with the data points in the viewing pane in the center of the tab. The R^2 value and equation of the line will be displayed in the upper right-hand corner of the tab. The user has the option of excluding the reference group from the correlation calculation by enabling the checkbox below the Calculate Correlation button.

As in all previous tabs, the Export Figure button will open the figure in the viewing pane for editing by MATLAB®, and the Export Data button will provide enable saving the tab analysis data as a .CSV file.

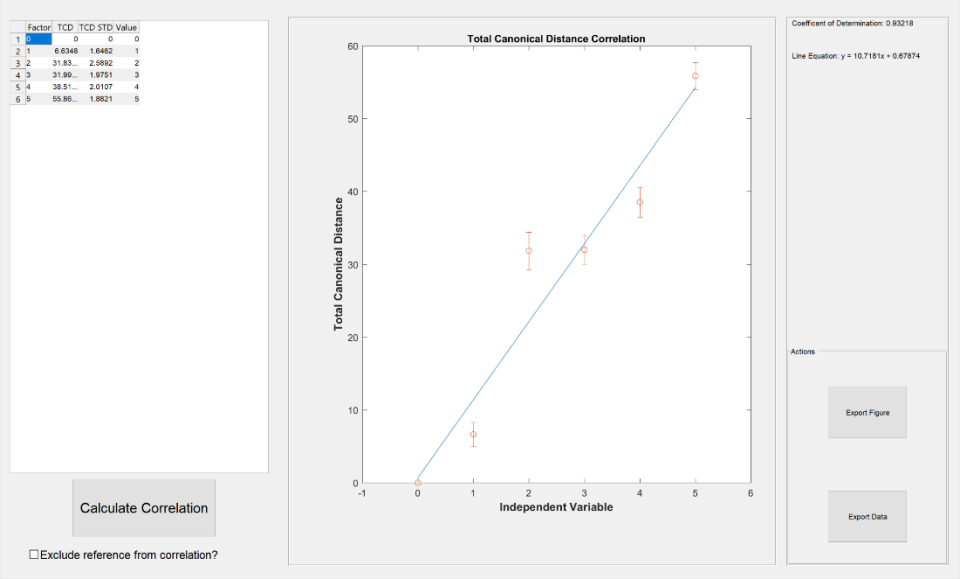

Figure S7. The TCD module.

Recommendations for Data Collection and Analysis

The following section includes recommendations to improve the effectiveness of analysis within the Rametrix™ LITE Toolbox.

Sample Collection

The most reliable data analysis will come from a small number of scans each of a large range of samples. Multiple scans of the same sample can either be used to compile an average spectrum of the sample, or be carried through analysis to DAPC. Sample size can be anywhere from only a dozen to several hundred or thousands of spectra. If the baselining calculations are taking too long, averaging of spectra is encouraged.

Spectrum names should contain multiple factors separated by underscores. Each factor represents a subcategory in a category denoted by the position of the factor in the spectrum file name. The first factors in every spectrum can be evaluated in DAPC together as *a priori* groups, as with the second factors in every spectrum, and so on. Every spectrum loaded in to the Rametrix™ LITE Toolbox must have the same number of factors in their file names. Factors should be selected based on the type of data analysis desired, bearing in mind that the arrangement of the groups calculated by DAPC can provide insight into the relationships between those groups.

Evaluation of scatter plots of Raman spectra provided by PCA and DAPC must be based on a solid understanding of the mathematical techniques to avoid erroneous interpretation. Raman spectra that are too similar to one another cannot be reliably interpreted via PCA. DAPC uses PCs as uncorrelated variables of lesser number than samples taken while still retaining as

much information as possible, but providing too much information results in overfitting group assignments, reducing the within-group variation to 0 and obfuscating potentially important inter-group relationships. Providing only one Raman spectrum per group category does not allow reliable interpretation via DAPC. For the best data, repeated trials are recommended where spectra come from different samples within the same category, as opposed to multiple Raman spectra of the same sample. As with all analytical techniques, increased sample size increases data resolution.

Selection of Goldindec Parameters

The algorithm used for baselining spectra requires two inputs specified by the user: baseline polynomial order and estimated peak ratio. The baseline polynomial order will modify the overall shape of the baseline. Look for the baseline for each spectrum in the 'Show Baseline' option in the Explore Tab to match the overall shape of the suspected background noise. Increase the baseline polynomial order if the baseline fails to follow the curve of one or more of the large, low peaks of background. The Goldindec algorithm will not attempt to follow the curve of tall peaks that are representative of sample features. Decrease the baseline polynomial order if the baseline has more peaks than the shape of the background.

The estimated peak ratio should be a value from 0.1 to 0.9, estimated in 0.1 increments. The Goldindec algorithm is robust enough to attain high accuracy when within 10% of the correct peak ratio. After overall baseline shape is attained as closely as possible with the baseline polynomial order, the estimated peak ratio should be increased if the baseline is too high, and decreased if the baseline is too low. Ultimately, the baseline should directly overlay the suspected background.

Selection of these parameters is best attained through trial and error via the use of the 'Show Baseline' option in the Explore module along the bottom of the screen. Figure S8 displays an example of baseline estimation and honing.

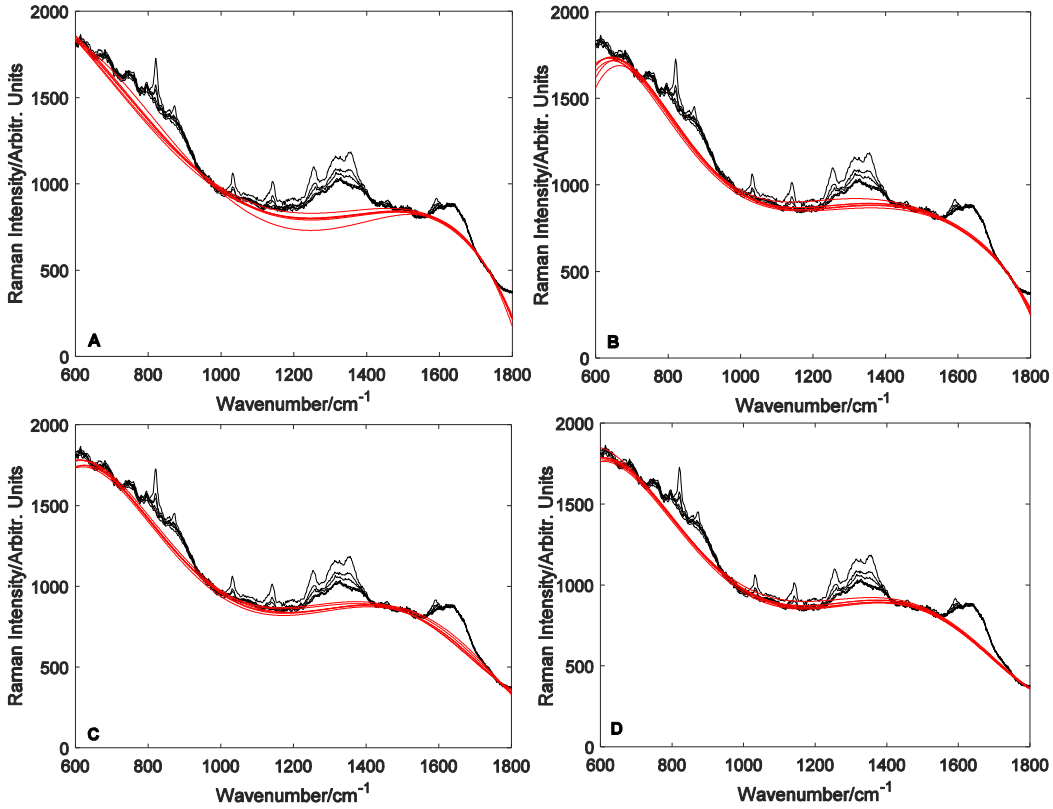

Figure S8. Baseline estimation example using the 2-nitrophenol concentration curve data, averaged. The best baseline polynomial order and estimated peak ratio for the data are 5 and 0.3. At first the default estimated peak ratio of 0.5 is used while the baseline polynomial order is adjusted to first A) 4 and then B) 6. Then the correct baseline polynomial order of C) 5 is applied, and D) the estimated peak ratio is adjusted to 0.3.

Visualizing Multiple Spectra in the Explore Module

The Explore module allows the visualization of a single or multiple spectra, as shown in Figure S8. Multiple spectra can be highlighted individually using the +control/command buttons in PC/Mac and selecting individual entries in the Explore module list of loaded spectra on the left. Or, a range of spectra can be selected using the +shift button and selecting the spectra at the ends of the range to view. We have noticed that selecting multiple spectra through hold-click (i.e., drag and drop) selecting does not work properly in MATLAB® at this time.

Applying Different Baseline Settings to Spectra Sets to be Evaluated Together

To avoid cluttering the interface of the Explore module, only one set of baseline options can be applied to any set of loaded spectra. To be able to evaluate spectra in PCA and DAPC that have been corrected to different baselines, a work-around is suggested.

Load the sets of spectra in separate batches and apply to each the desired baseline options. Use the Export Data button to save each set of the baselined (and normalized) spectra

to .CSV files in a chosen folder. Once all preprocessing has been completed, load the baselined (and normalized) spectra into the RDA Toolbox together and keep the popup menu in the Explore module set to 'None' when performing PCA and DAPC. It is not possible to normalize spectra without first applying baseline calculations in the Rametrix™ LITE Toolbox, so if normalization is desired it is recommended to be carried out before saving the .CSV files.

The Effect of Variability Inclusion on DAPC Analysis

PCA provides at least n-1 principal components that each describe a portion of the total system variability. The first principal component will account for the greatest possible variances in a linear arrangement of the wavenumber intensities, and each subsequence principal component will represent consecutively less system variability.

The number of principal components used in the DAPC calculations must account for less than 100% of the total system variability, and thusly the maximum number of principal components that can be used is one less than the total number calculated. However, too many principal components can perfectly discriminate sampled individuals, eliminating any within-group variation to an unrealistic degree, demonstrated in Figure S9. Note how the scale of the axes grows more dramatic, and how the within-group variability diminishes as more system variability is provided to the DAPC calculation. There is currently no consensus on the best amount of total system variability to consider in DAPC analysis. The default of 90% is only meant to provide a useful starting point.

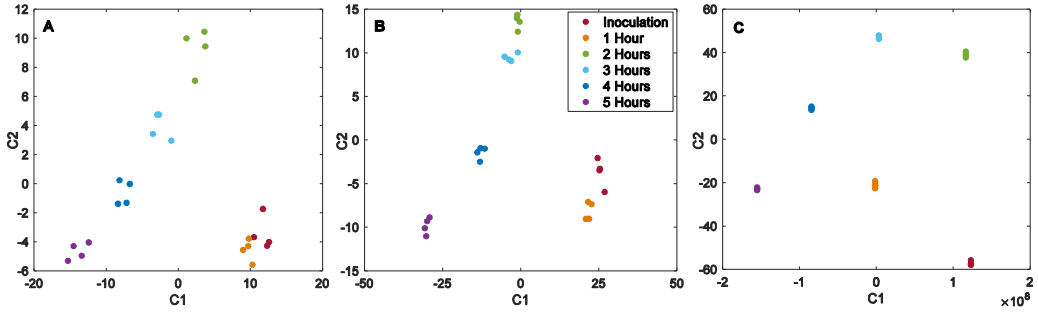

Figure S9. *E. coli* 10-β culture growth from inoculation to 5-hour incubation in 37°C at 200 RPM interpreted via DAPC using A) 90%, B) 95%, and C) 98% of the total system variability.