# SpectroChat: A windows executable graphical user interface for chemometrics analysis of spectroscopic data

## User manual

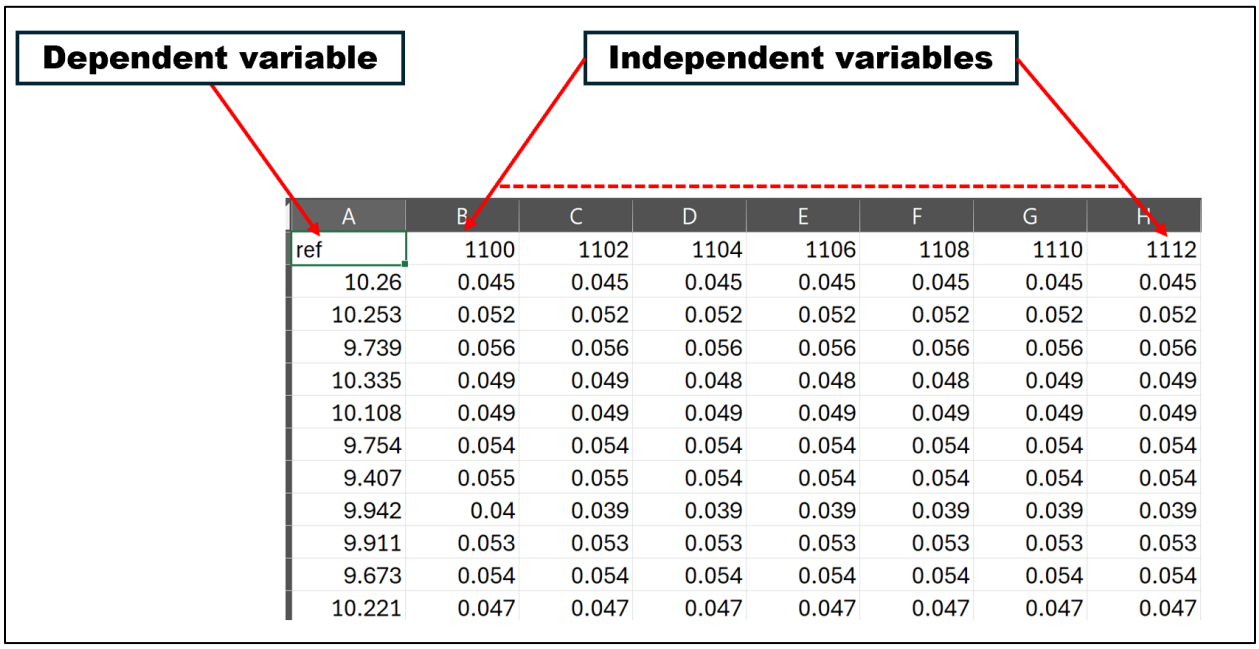
List of contents

## 1. Downloading and Installation

Go to https://github.com/SpectroChat/GUI/blob/master/SpectroChat.exe and click on the download button. After downloading the software, users can click on SpectroChat.exe to start the analysis.

## 2. Input data preparation

Like other software, SpectroChat software is heavily reliant on the proper formatting of the input data before the analysis is executed. SpectroChat requires users to prepare input data in the comma-separated values (CSV) format. Besides, the CSV file should contain a header, the dependent variable header should be named "**ref**" and the independent variables should be named in numerical values (integer/float) (depicted in Fig. I1). A sample of corn spectroscopic data is formatted and uploaded to https://github.com/SpectroChat/GUI/blob/master/sample_data.csv. The dataset is retrieved from http://www.eigenvector.com/data/Corn. Users can follow this "sample_data.csv" format to prepare their own dataset for analysis using SpectroChat.




| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| ref | 1100 | 1102 | 1104 | 1106 | 1108 | 1110 | 1112 |
| 10.26 | 0.045 | 0.045 | 0.045 | 0.045 | 0.045 | 0.045 | 0.045 |
| 10.253 | 0.052 | 0.052 | 0.052 | 0.052 | 0.052 | 0.052 | 0.052 |
| 9.739 | 0.056 | 0.056 | 0.056 | 0.056 | 0.056 | 0.056 | 0.056 |
| 10.335 | 0.049 | 0.049 | 0.048 | 0.048 | 0.048 | 0.049 | 0.049 |
| 10.108 | 0.049 | 0.049 | 0.049 | 0.049 | 0.049 | 0.049 | 0.049 |
| 9.754 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 |
| 9.407 | 0.055 | 0.055 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 |
| 9.942 | 0.04 | 0.039 | 0.039 | 0.039 | 0.039 | 0.039 | 0.039 |
| 9.911 | 0.053 | 0.053 | 0.053 | 0.053 | 0.053 | 0.053 | 0.053 |
| 9.673 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 |
| 10.221 | 0.047 | 0.047 | 0.047 | 0.047 | 0.047 | 0.047 | 0.047 |

Fig. I1: Formatting input CSV data for SpectroChat. The dependent column should be named "**ref**" (**case sensitive and without any space**) in the header, while the independent column should be named with numerical values.

## 3. Data partitioning

After ensuring the proper formatting of data, users can select the data from the software prompt for further analysis. The consecutive series of actions for partitioning the data is illustrated in Fig. I2(a-h). Firstly, users have to click on "Data partitioning" (Fig. I2a). Next, users should click on "Select File" (Fig. I2b). Then, users will be directed to locate the formatted input file, and users have to select the file (Fig. I2c). After that, users have to input the number of training samples to be in the partitioned training file (the rest of the samples will be in the test file) and press the "**Enter**" key on the keyboard (Fig. I2d). Users can use the dropdown menu (indicated in Fig. I2e) to select different data partitioning methods and then click the "Split" button. After successful steps, users will be shown the partitioning method performed and the location of the generated partitioned files at the bottom of the software (Fig. I2f). Users can go to the "Partitioned data" folder in the same software directory and locate the generated train and test files inside it (Fig. I2g-h).

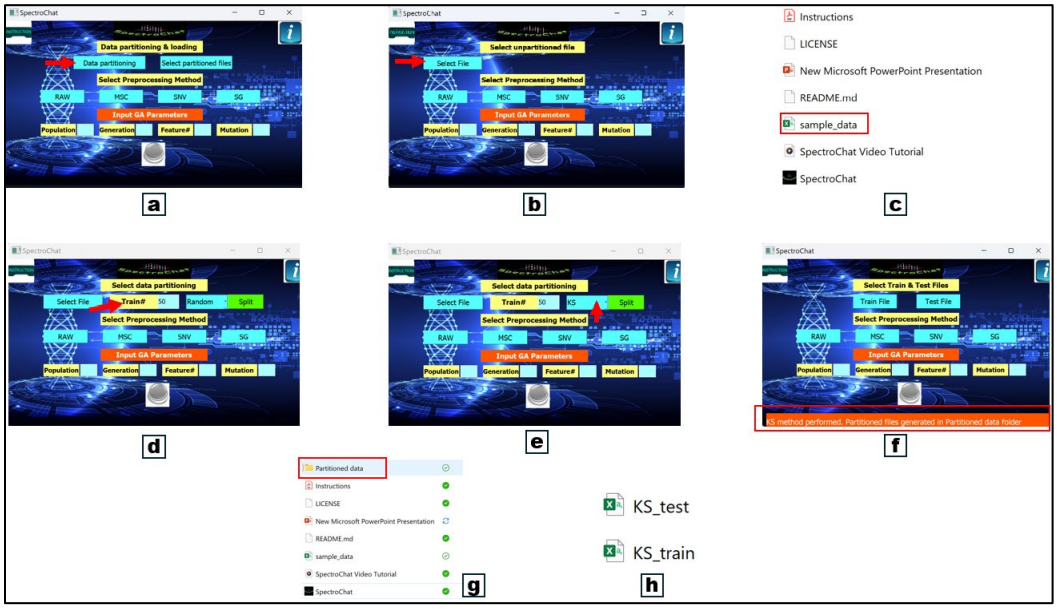



Fig. I2: Data partitioning steps in SpectroChat. (a) Clicking the "Data partitioning" button, (b) Selecting the data file, (c) Locating the data file to be partitioned, (d) Inputting the number of training samples and pressing the "**Enter**" key, (e) Choosing the partitioning method from the dropdown menu, (f) Successful execution of data partitioning, (g, h) Locating the partitioned files in the same directory of the SpectroChat.

## 4. Loading the train and test files

After the successful execution of the data partitioning users can load the train and test files. The series of consecutive easy steps are shown in Fig. I3(a-e). After loading the datasets, users will be displayed "Files Selected," and users will be urged to go to the next stage of the analysis. Users can also choose their own partitioned training and test files by skipping the "Data partitioning" option and clicking the "Select partitioned files" option (refer to Fig. I2a) after the initial execution in a similar way. In that case, users should check the data formatting as in section 2 to ensure proper execution.

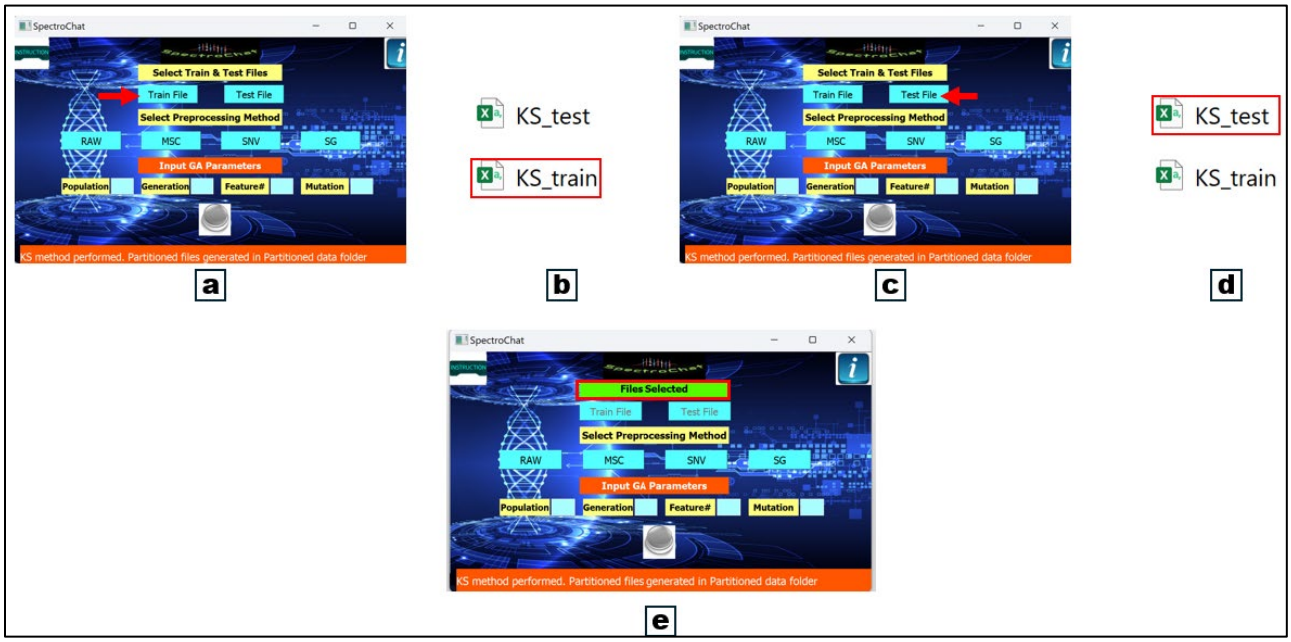



Fig. I3: Loading train and test files. (a-b) Selecting Train File, (c-d) Selecting Test File, (e) Window after file selection.

## 5. Selecting pre-processing method

SpectroChat offers users four preprocessing methods: RAW (original spectra), MSC, SNV, and SG (Fig. I4a). After successfully loading the training and test files, users can choose either of these methods. After selecting one of these methods, plots showing the training and test spectra will be displayed (Fig. I4b-c). These plots can be edited according to the experiment (titles, label changes, saving in different formats, etc. as depicted in Fig. I4d). For the SG preprocessing methods, users can input the parameters der, poly, and gap and then click on "RUN SG" (steps depicted in Fig. I5a-d). Users should be well informed about the SG preprocessing method for the choice of the parameters.

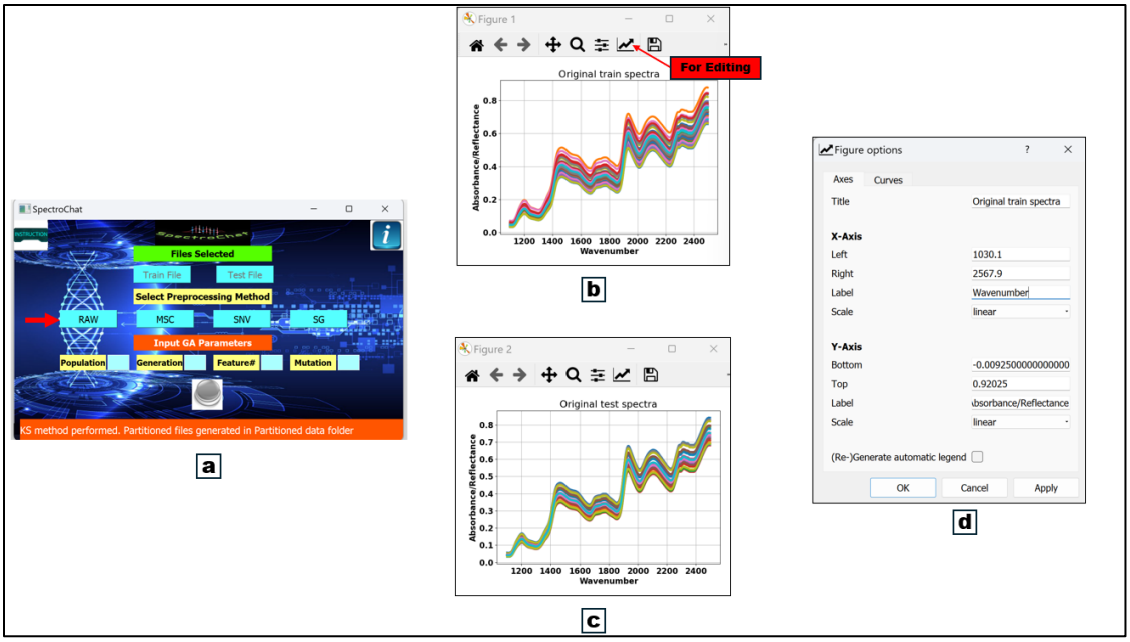



Fig. I4: Pre-processing steps. (a) Selecting preprocessing type, (b) Preprocessed train spectra, (c) Preprocessed test spectra, (d) Options for editing plots.

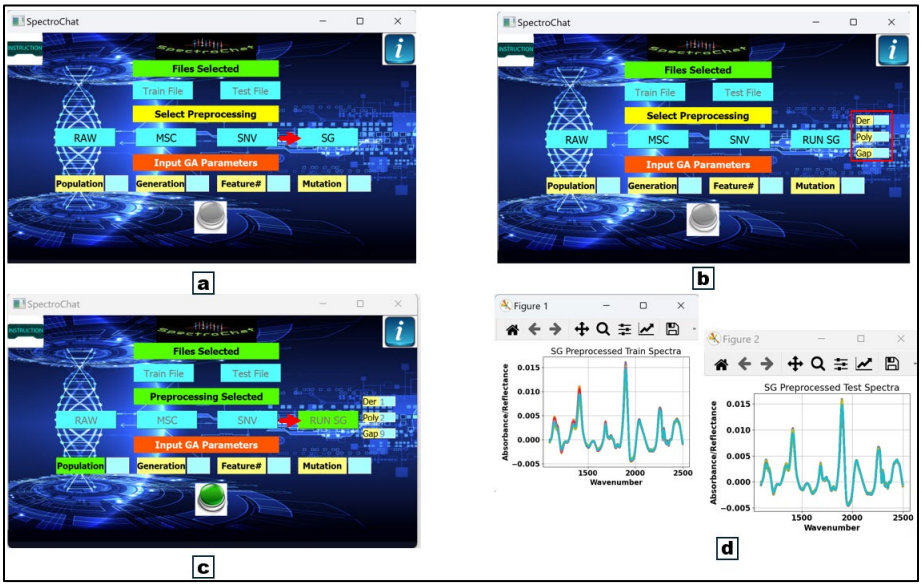



Fig. I5: SG Pre-processing steps. (a) Selecting SG preprocessing type, (b)Inputting parameters according to the data and analysis needs, (c) Clicking “RUN SG” for calculation, (d) SG pre-processed plots.

## 6. Regression model using all the features

After selecting the pre-processing method users can run the regression model using all the features. Firstly, users have to click on the button without giving any input to the GA parameters (Fig. I6a). Next users have to click on the "SHOW RESULTS" button. Then, users will be directed to the output window containing the option to use PLSR and MLR models (Fig. I6c).

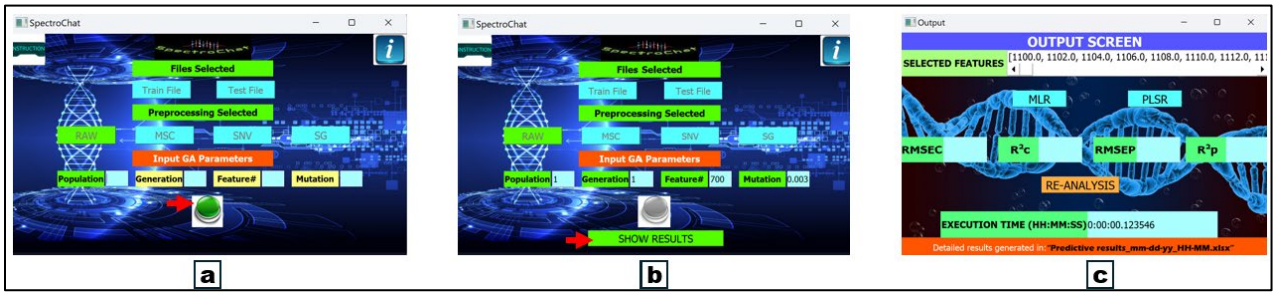



Fig. I6: Steps for modelling with all features. (a) Pressing the button directly without giving any input to the GA parameters, (b) Clicking the "SHOW RESULTS" button, (c) Output window containing modelling and result screen.

### 6.1 PLSR model using all features

PLSR is the most commonly used regression method in spectral analysis, especially with full features. In the output window, users can click on the PLSR button (Fig. I7a). After clicking the button, a plot containing the suggested LV by the Haland-Thomas criterion will appear to help users choose the desired LV (Fig. I7b).

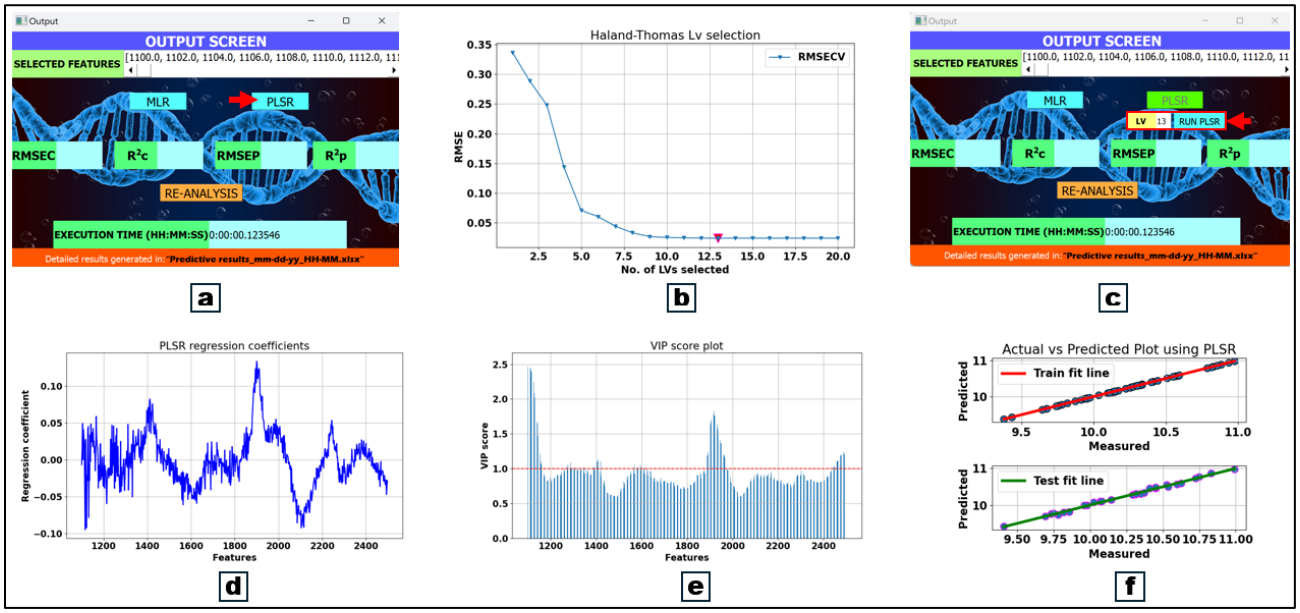


Fig. I7: PLSR modelling using SpectroChat. (a) Clicking the "PLSR" button, (b) Optimized LVs suggested by SpectroChat using the Haland-Thomas criterion, (c) Inputting LV and clicking on "RUN PLSR" button, (d) Regression coefficient plot, (e) VIP plot, (f) Fitting line.

Next, users have to input the LV and click on the "RUN PLSR" button as depicted in Fig. I7c. After clicking the button, users will have a PLSR regression coefficient plot (Fig. I7d), VIP plot (Fig. I7e), and fitting line plot (Fig. I7f). All these plots can be helpful for further diagnosing the analysis and modelling decisions. Besides, the plots are also customizable like the plots shown in Fig. I4.

After the successful analysis, users can see the regression results on the screen (Fig. I8a). Besides, results are also saved in a .xlsx file in the same directory with the analysis time as the name (Fig. I8b-c).

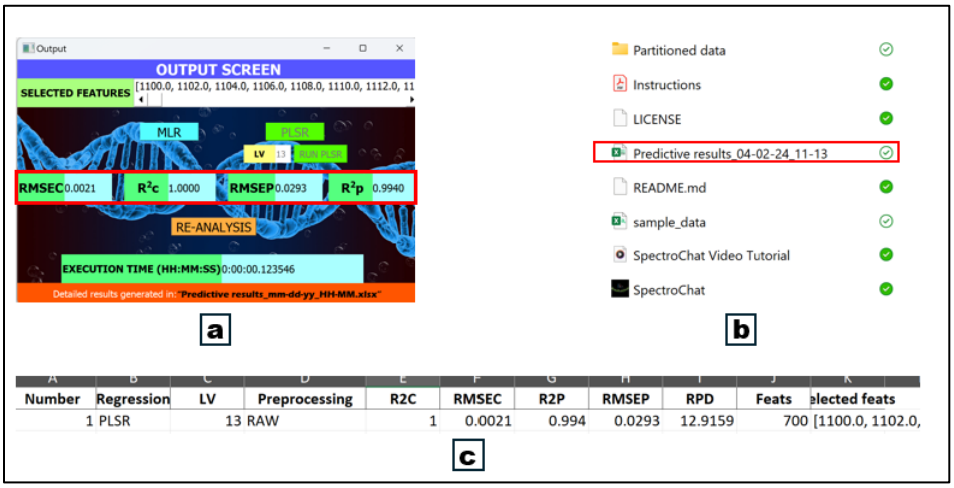



Fig. I7: Output results of the model using SpectroChat. (a) Regression results on the output screen, (b) The .xlsx file saved in the same directory, (c) Detailed result inside the .xlsx file.

### 6.2 MLR model using all features

Users can also perform MLR model on the data. Though MLR is commonly performed with feature selection in spectroscopic analyses. The visuals of MLR modeling is displayed in section 7.

## 7. Feature selection using GA

SpectroChat allows users to select some important features using the genetic algorithm (GA). After the preprocessing steps (section 5), users can input the four parameters: population, generation, number of features, and mutation rate (Fig. I9a). Then users can click on the  button, followed by the "SHOW RESULTS" button as depicted in Fig. I6. After successful feature selection steps, users will be displayed a plot with the selected features on the spectrum (Fig. I9d). Also the output window now will contain the list of selected features (Fig. I9e).

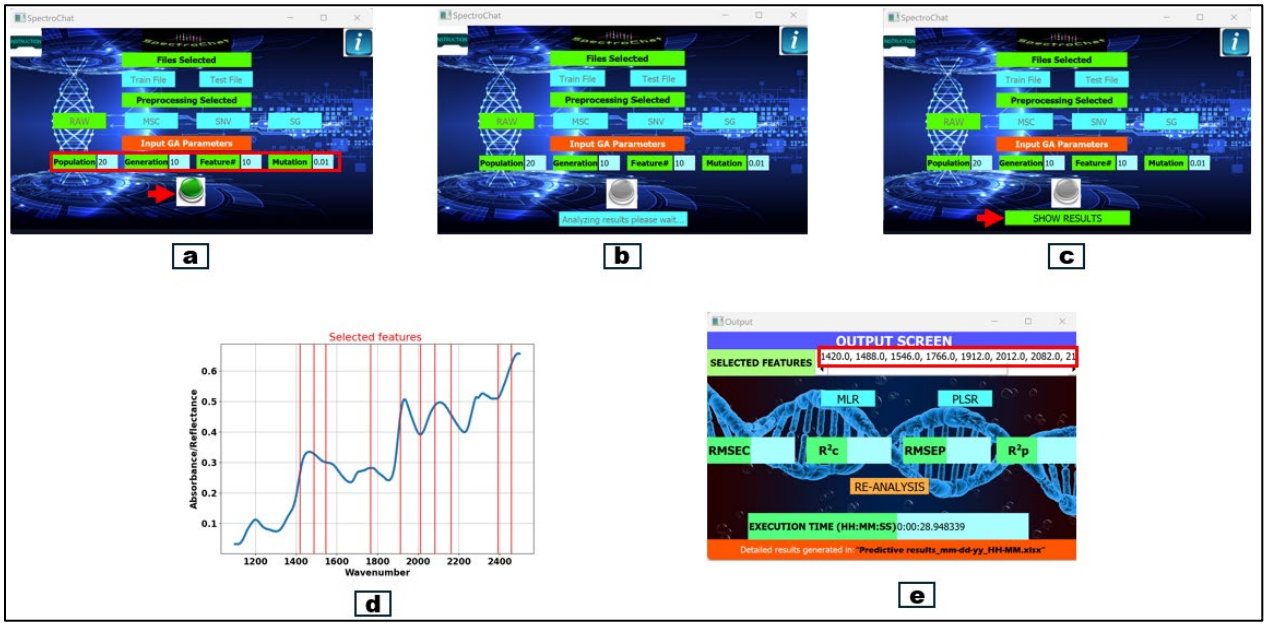



Fig. I9: Feature selection by GA using SpectroChat. (a) Inputting GA parameters, (b) Waiting window for feature selection, (c) Clicking on the "SHOW RESULTS" button, (d) Selected features shown on a spectrum, (e) Output window containing list of selected features.

### 7.1 Modelling on the selected features

As shown in section 6.1, modelling can also be performed on the selected features using both MLR and PLSR. Fig. I10a-c contains the steps to perform the MLR modelling on the selected features. After the modelling the results will also be saved in a .xlsx file as shown in Fig. I7.

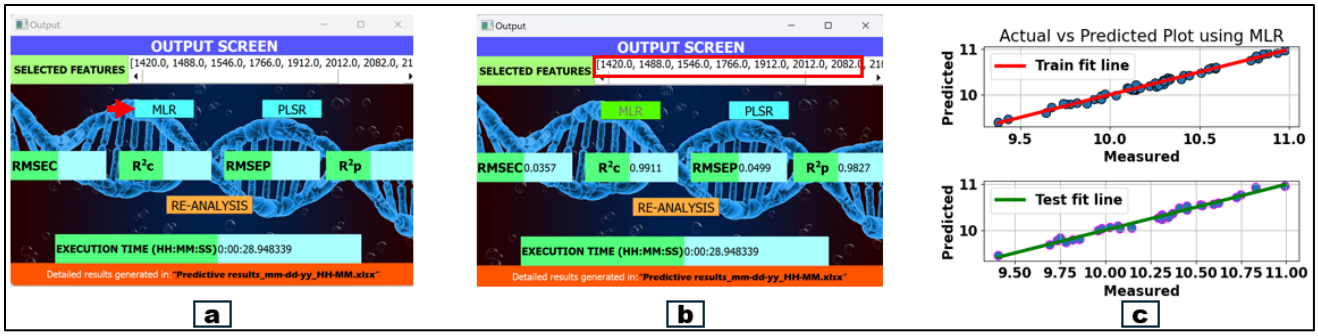



Fig. I10:MLR on the selected features. (a) Clicking on the "MLR" button, (b) Window containing list of features and regression results, (c) Fit line using MLR.

## 8. Reanalysis

Sometimes, users might have to reanalyze the same data. SpectroChat provides options to reanalyze by selecting different preprocessing methods and selecting a different set of features. Users have to click on the "RE-ANALYSIS" button on the output window and start from section 5 (Fig. I11).

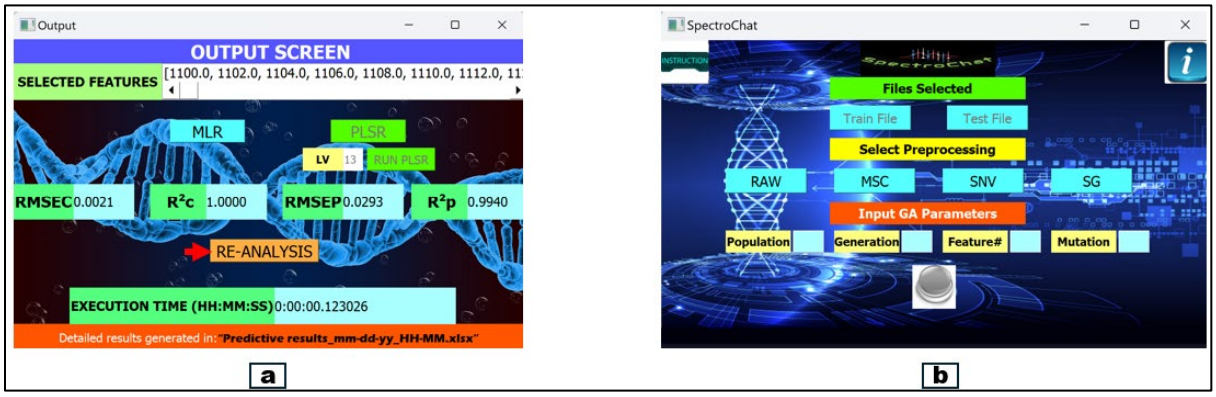



Fig. I11: (a) Click on the "RE-ANALYSIS" button, (b) Start from preprocessing steps (Section 5).

## 9. Instruction and info page

Apart from the analysis, SpectroChat offers two more windows containing brief instructions and info. Fig. I12 demonstrates glimpse of the windows.

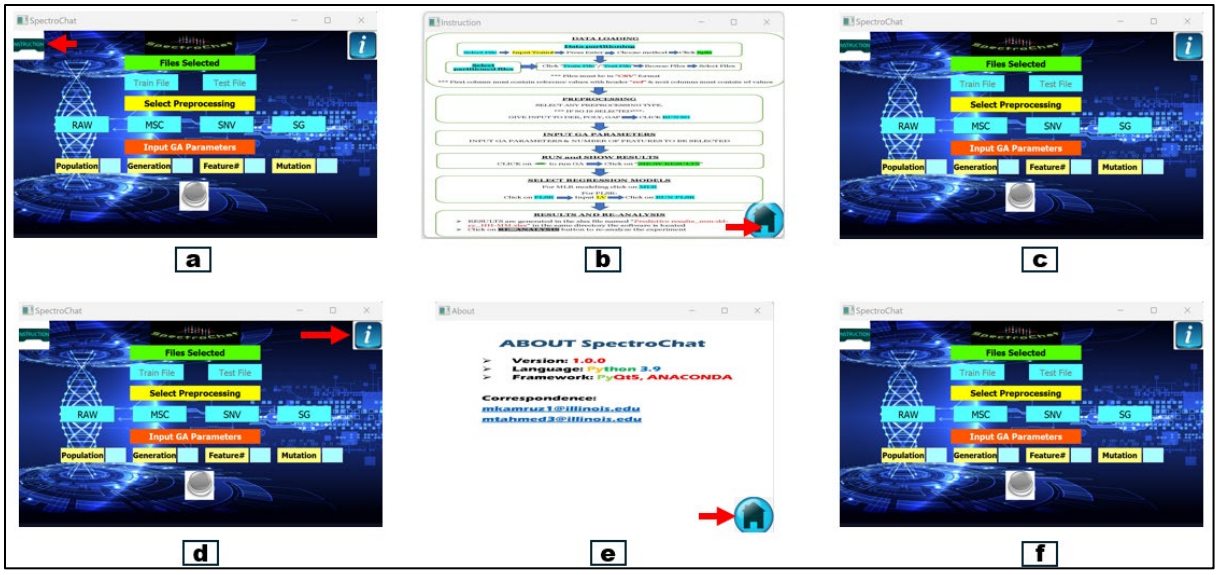



Fig. I12: Switching between the Instruction and about windows from and to the main window.