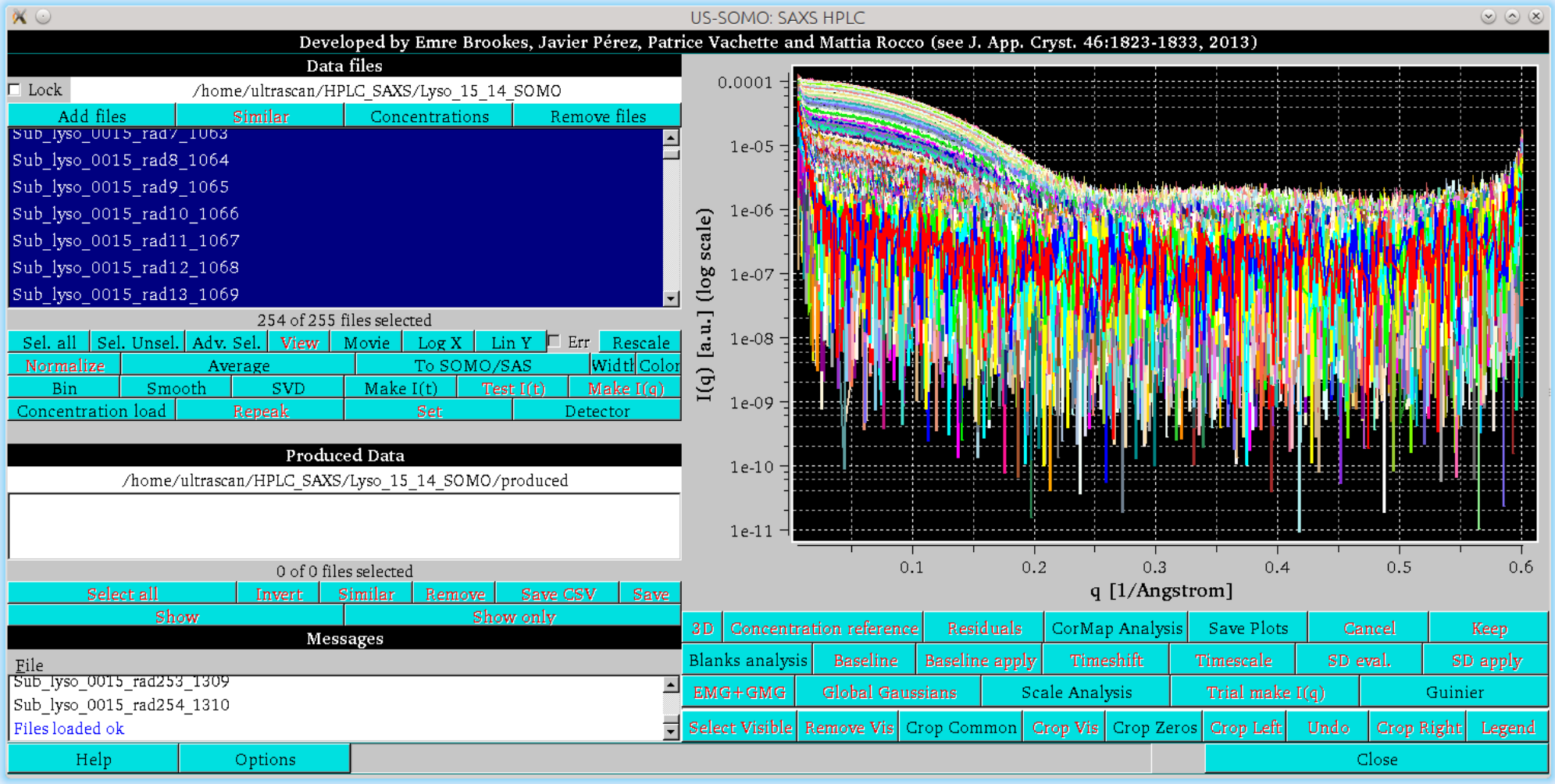
Table of Contents

SOMO HPLC-SAXS Module:

Last updated: January 2018

NOTICE: this module is being developed by E. Brookes, J. Pérez, P. Vachette, and M. Rocco.
Portions of this help file are taken from the Supplementary Materials of Brookes et al., "Fibrinogen species as resolved by HPLC-SAXS data processing within the UltraScan SOlution MOdeler (US-SOMO) enhanced SAS module", J. Appl. Cryst. 46:1823-1833 (2013), and from Brookes et al. "US-SOMO HPLC-SAXS Module: Dealing with Capillary Fouling, and Extraction of Pure Component Patterns from Poorly Resolved SEC-SAXS Data", J. Appl. Cryst. 49:1827-1841, 2016. Some recent improvements are discussed in Brookes and Rocco, Eur. Biophys. J., 2018, submitted



This **US-SOMO** module was conceived for the analysis of HPLC-SAXS data. In the image above, the main panel of the ***HPLC-SAXS*** module is shown. The buttons with the black labels are the ones currently active, the ones with the red labels become active when allowed by the processing/visualization stage. The graphics panel shows a collection of HPLC-SAXS $\log_{10}[I(q)]$ vs. q SAXS data frames (points with 0 or negative values are automatically omitted from the visualization only) for a chicken egg-white lysozyme chromatographic separation on a Agilent BioSec-3 (3 µm particle size, 300 Å pore-size) 4.6 × 300 mm column, eluted with Hepes 50 mM, NaCl 100 mM, pH 7. Note the permanent upturn at very small q-values, due to biological material aggregated by the intense X-ray beam on the capillary cell walls under these far from optimal experimental conditions. While this kind of problem should be (and has been) preferentially dealt with at the experimental level, we use this dataset to demonstrate the potential for correcting data still presenting such an issue.

The left side of the window is divided in three sections, labeled "***Data files***", "***Produced Data***", and "***Messages***". By clicking on these labels, the corresponding panel below each label will disappear, allowing for an expansion of the remaining other panel(s). If every panel is made to disappear, the main graph will expand to cover the full size of the ***HPLC-SAXS*** window. By clicking again of the labels, the corresponding panels will be restored.

On the top left panel (***Data files***) there are four buttons:

The ***Add files*** button is used to load data into the module. An operating directory can be pre-selected by clicking on the path shown above it, and navigating in the file system (selecting the *Lock* checkbox will fix that directory). The file format for SAXS data recognized by the **US-SOMO** ***HPLC-SAXS*** module consist of .dat files with two or three TAB- or space-separated columns containing the q, $I(q)$, and optionally their associated standard deviation (SD) values, respectively. Each frame number (or time value) must be present somewhere in the filename with a common prefix and suffix. For example, data1saxs.dat, data2saxs.dat, data3saxs.dat will be recognized as frames 1,2,3, where "data" and "saxs" can be replaced by any common sequence of characters. Consequently, 1.dat, 2.dat, 3.dat would be acceptable, but abc1.dat, qrs2.dat, xyz3.dat would not, because the prefix characters are not common. Furthermore, the loader will also arrange the data files sequentially, in increasing frame number (or time value) order. Concentration-related data should be instead uploaded using the ***Concentration load*** button (see below). $I(q)$ vs. q and concentration data frames are automatically recognized and the labels on the x- and y-axes are then properly set.

Similar will select files with similar names and allow manual pattern matching entry if no new similar files are selected.

Concentration will show every file listed together with their associated concentration (mg/ml), if appropriate and properly set (see below). Concentrations can also be entered and modified manually. They can be used to normalize the $I(q)$ vs. q data (see below). Loaded files can be displayed on the graphics panel by individually clicking on them (shift-click will select a contiguous series, ctrl-click allows multiple irregularly spaced selections). Produced data will also show up in this panel with associated putative filenames.

Remove files will discard previously selected files (see below); if the files were produced by the module, and were not previously saved, a warning window will pop-up, allowing to proceed or to stop removing the selected items.

Several buttons are available in the panel below the loaded files window:

Sel. all will select all files.

Sel. Unsel. will allow toggling the selection between selected files and everything else not currently selected.

Adv. Sel. will open up a panel in which several selection options can be utilized (see here).

View, active when up to ten datasets are selected, will show them in text format.

Movie: Pressing this button will open a pop-up window with the commands allowing to view in the main graphics window of the **US-SOMO** ***HPLC-SAXS*** module a series of selected data files in a movie-like manner, and to optionally save each frame as an image for real movie-making operations (see here).

The ***Log X*** (***Lin X***) and ***Log Y*** (***Lin Y***) buttons allow to toggle between linear and $\log_{10}$ scaling of the data on the x- and y-axes, respectively (if zero or negative values are present, they will be temporarily removed when the scale is set to $\log_{10}$ mode, as they cannot be shown on the display in this mode). The buttons will change their respective label once pressed, to underscore what is the action currently

available.

Selecting the ***Err*** checkbox, active when up to 10 files are selected, will switch their representation from the dots connected with a line mode, to symbols (diamonds) with their associated SDs represented as error bars mode.

Rescale adjusts the X-Y axes on the graphics window to maximize the display of selected datasets (no effect on the data themselves).

Normalize will divide the $I(q)$ data by the stored/entered concentrations.

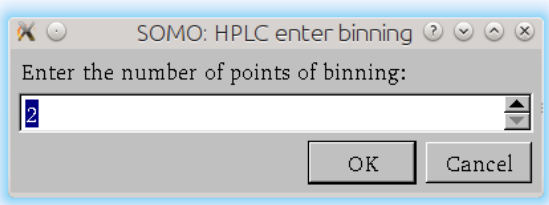
Average will produce an average with propagated SDs of selected data. The $I(q)$ values from each frame will be averaged, and then a scaling factor will be determined for each frame against the average resulting frame. The scaling factor for each original frame multiplies the frames's SD. The average intensity SD's are computed as the square root of the sum of the squares of each curves scaled SD's, and this is divided by the number of curves. The resulting datset filename will contain the number of frames averaged, and the initial and final frame numbers, followed by "_avg".

To SOMO/SAS will transfer selected datasets back into the **US-SOMO *SAS*** panel.

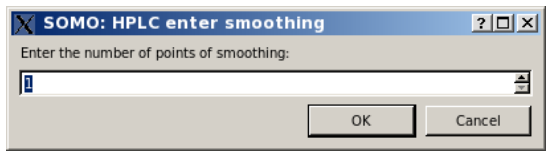
Each time the ***Width*** button is pressed, it increments the data line (or symbol) size of the plots, until it goes back to the initial value.

Color shifts the colors used in the graphics window; the operation can be repeated until a better contrast with the background is achieved. Note that the background color can be changed by right-clicking on the plot borders, which will open up a pop-up dialogue panel where all plot characteristics can be modified.

Bin allows averaging adjacent points in $I(q)$ datasets, starting with the first point in the file and using a binning size defined in a pop-up dialogue:



Smooth performs a regularization of selected data using a moving window, whose dimension is defined in a pop-up menu (shown below), using a Gaussian smoothing kernel of 2 n+1 points.



SVD opens a pop-up window were a single-value decomposition analysis (e.g., Williamson et al., Biophys J. 94, 4906-4923, 2008) can be performed on the selected data (see here). Important: the data must be all on the same grid; if not, a warning message will appear in the bottom left ***Messages*** window: "SVD: curves must be on the same grid, try 'Crop Common' first" (see below for the use of the ***Crop Common*** button).

Make I(t) is one of the crucial operations in the ***HPLC-SAXS*** module. It allows to generate a series of "chromatograms" ($I(t)$ vs. t, where t can be real elution time or frame number) for each q-value present in the original data files (see below). A test could be automatically performed each time an $I(q)$ vs. q dataset is converted into an $I(t)$ vs. t dataset to ascertain if any $I(t)$ vs. t "chromatogram" contain useful data, on the basis of a comparison between the signal and its associated SDs, by selecting its relative checkbox and the SD factor in the ***Options*** menu accessible from the button provided at the bottom of this window (see here).

Test I(t) Checks the $I(t)$ vs. q selected curves to see if any fail the negative region test as described above.

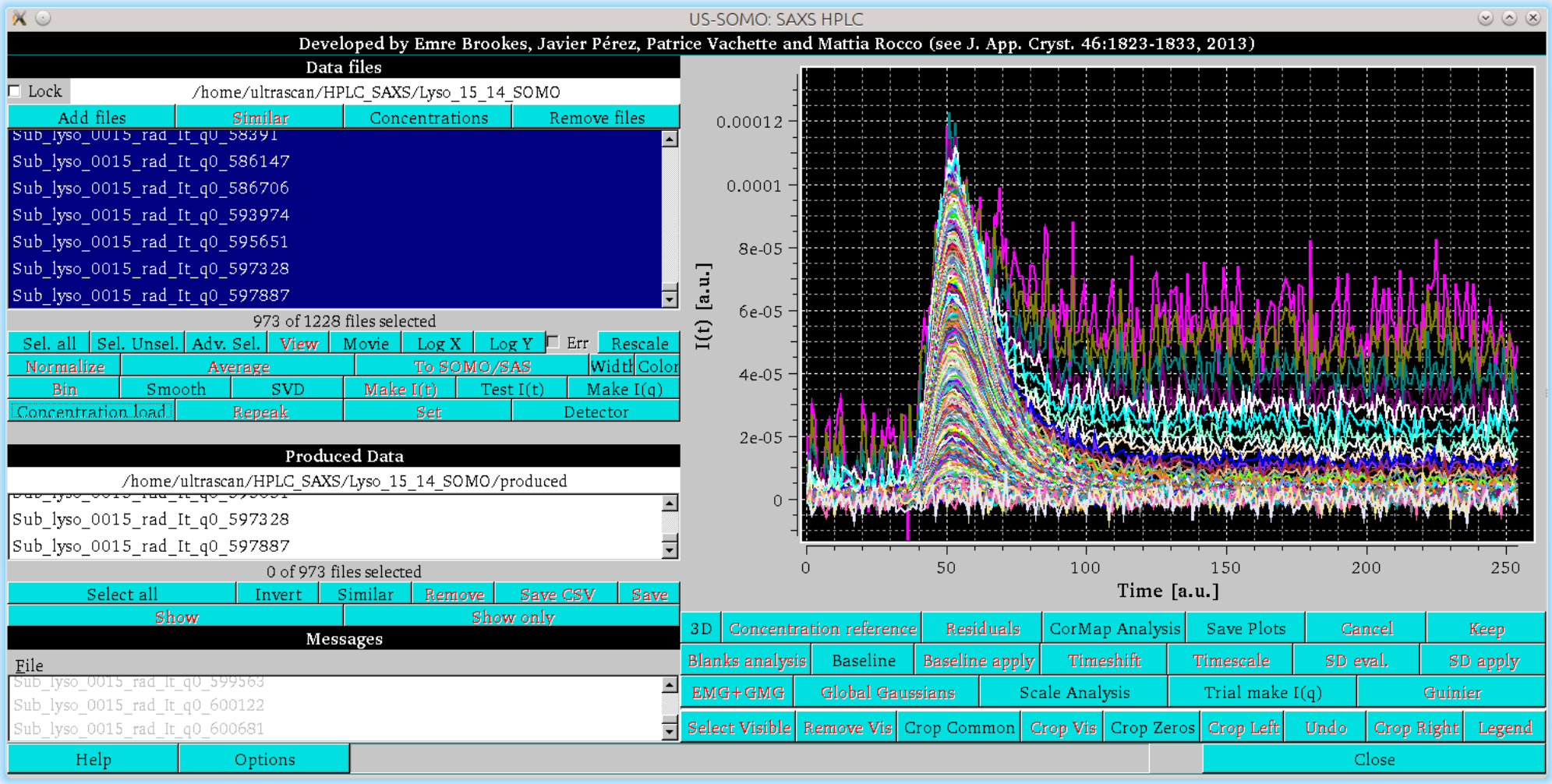
Make I(q) is the other crucial operation in the ***HPLC-SAXS*** module. It allows to re-generate $I(q)$ vs. q files for each frame after data treatment in frame- (or time-) space.

Concentration load is used to upload any chromatographic data files containing a concentration-related elution profile, such as those produced by UV-VIS absorption or refractive index detectors (the program will then internally keep track of such datasets, distinguishing them from SAXS datasets). By default, the program will look for "*.txt" files, but the choice could be expanded to other extensions in the file upload dialogue. The currently recognized format for concentration data is similar to the SAXS data format with the addition of the string "Frame data" in any place on the first line. The two or three columns of data are the frame number, concentration-related data, and optionally an associated SD value.

Repeak is used to effectively scale data (usually a concentration-related chromatogram) on the y-axis to a pre-set target (usually a low-q, high-intensity $I(t)$ vs. t chromatogram), selectable in a pop-up window among the data subjected to this operation (this affects the data, a new file is generated with "rp" and the scaling factor added at the end of the filename). See more below on this subject.

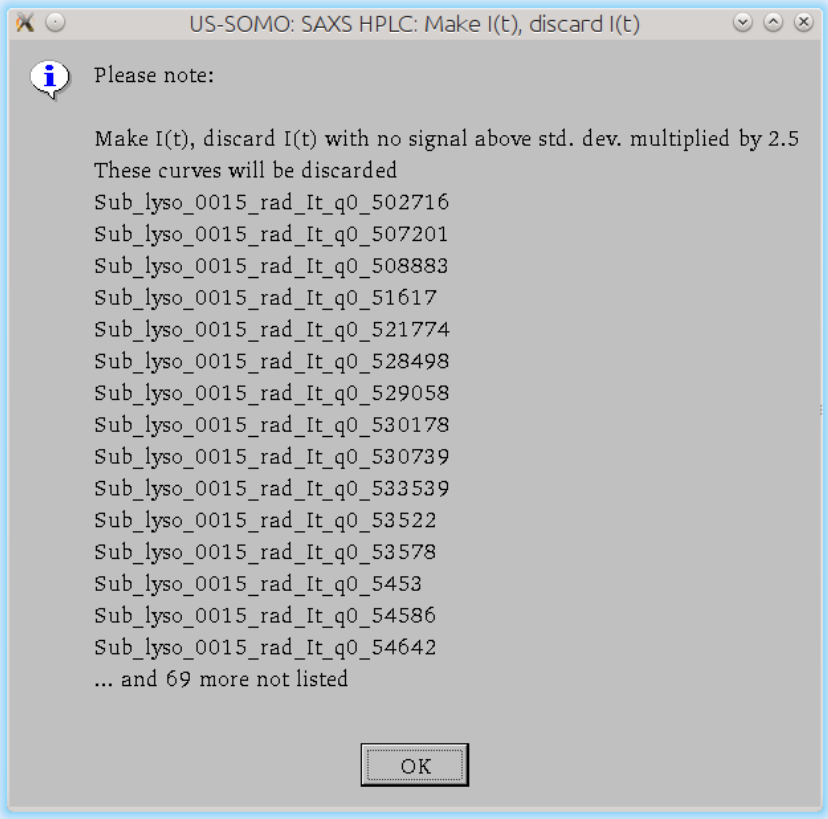
Set will set an already uploaded and currently selected file containing the UV or refractive index profile vs. time or frame number as the source of the concentration-dependent signal.

Detector will allow to select the type of detector and to enter its calibration constant in a pop-up window (see here).

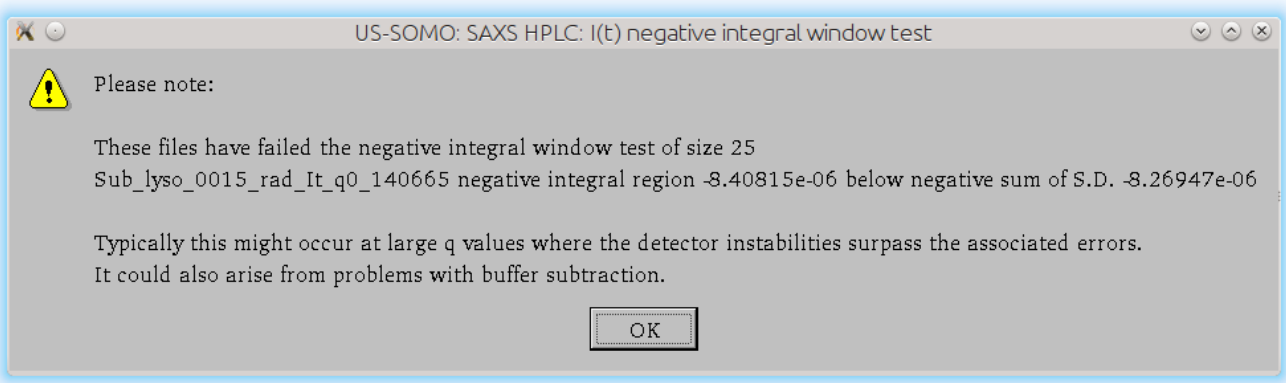


Since a typical HPLC-SAXS experiment produces a series of $I(q)$ vs. q data collected at some time interval ("frames"), they can be inserted in a 2D matrix where each line corresponds to a frame number (or

time value) and the columns contain the intensities *I(q)* and their associated SDs at the various scattering angles *q*. It is then a simple operation to generate another matrix where the lines correspond to the *q*-values and each column contains the intensities *I(t)* (and their associated SDs) corresponding to each frame number (or time value). A new data set consisting of *I(t)* vs. *t* "chromatograms" for each *q*-value can then be generated.
In the image above, the original *I(q)* vs. *q* data shown in the first image of this Help section have been transformed to *I(t)* vs. *t* data by pressing the ***Make I(t)*** button after selecting all files. The *I(t)* vs. *t* data are automatically displayed after the conversion, and the *q* values are now part of the resulting filenames. Since the ***On Make I(t), discard I(t) with no signal above st. dev. multiplied by*** "2.5" was selected in the ***Options*** menu (see here), the following *Warning message* appeared:



In addition, a test is automatically performed to identify regions within a sliding window (of 25 frames in this case) where the sum of the intensity is less than the negative of the sum of the corresponding SD values over the window. Regions with negative values could cause problems with the integral baseline subtraction procedure (see more below). This test identified just a single *I(t)* vs. *t* chromatogram failing it, as shown in a pop-up window:



Some cropping operations (see below) can be also performed to remove very noisy low-*q* datasets, such as the first three *q* values displayed in the Figure above (magenta, olive and greenblue) and/or to truncate the datasets if necessary. All operations are recorded in the bottom left panel.

The file names of produced data are shown in the ***Produced Data*** panel to the centre-left, and can be selected and saved to files using the appropriate buttons below it.

Select all will select all files in this panel.

Invert will allow toggling the selection between selected files and everything else not currently selected.

Similar will search for similar file names after selecting a single file in this panel.

Remove will discard the selected files.

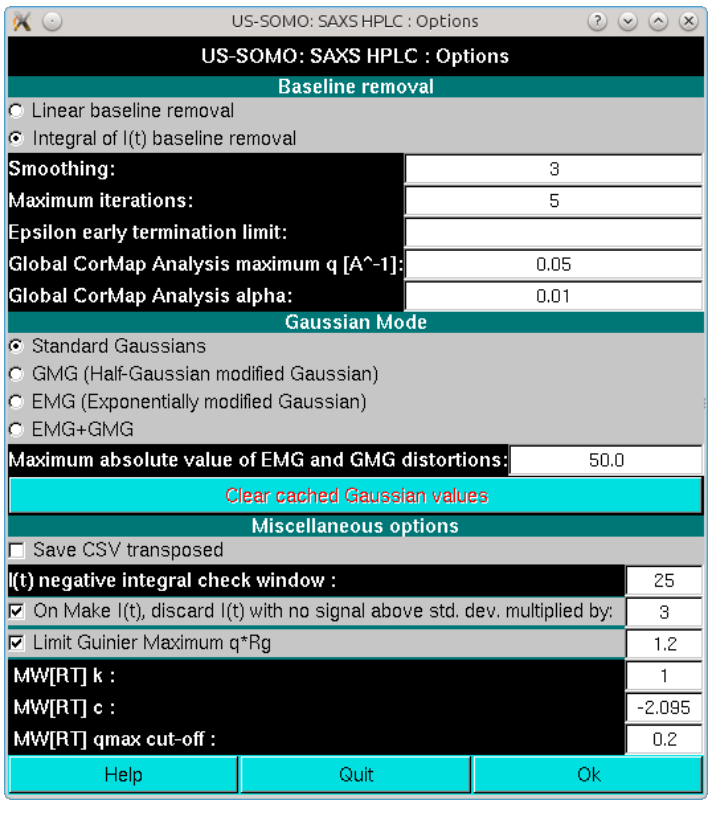
Two types of files can be produced, csv-style (***Save CSV***) or regular 3-columns .dat files (***Save***).

Show will add the selected file(s) among those produced to the ones already displayed in the graphics window.

Show only will show only the selected file(s) among those produced in the graphics window.

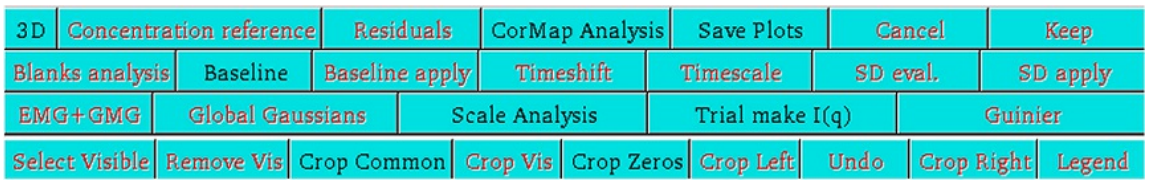
In the ***Messages*** area, the operations performed are tracked, and computed parameters are shown. The display can be copied or cleared from the *File* pull-down menu.

The last line of the left-side panels contains the ***Help*** and ***Options*** buttons. On pressing the latter, a pop-up panel will be shown:



See here for a description of this module.

Below the **US-SOMO *HPLC-SAXS*** module graphics panel there are a series of buttons for performing several operations on the files displayed, some of which will become available only when multiple files are selected, or a region of the graph is zoomed, while others will become available only when single files are selected:



When a part of the graph is selected using the mouse/left button, the buttons in the bottom line become all available (only ***Crop Zeros*** and ***Crop Common*** are available when files are just displayed after selection).

- ***Select Visible*** will select the files shown in the graphics window, which can be zoomed using the mouse (left click & drag). For instance, this is a practical way of selecting only a few files, by zooming on a region where only them are present.
- ***Remove Vis(ible)*** will instead remove the files shown in the graphics window.
- ***Crop Common*** will crop all selected files so that they have identical x-axis values by dropping points outside of the union of all selected file's x-axis values.
- ***Crop Vis(ible)*** will remove what is shown in the graphics window. For instance, this is practical way to trim the data on the *x*-axis.
- ***Crop Zeros*** will remove data having zero or negative values in the intensity columns (discouraged: a warning panel will pop-up, asking the user whether she/he really wants to proceed, since experimental zero or negative values have statistical significance).
- ***Crop Left*** and ***Crop Right*** will remove one point on the left or right of the selected data, respectively.
- ***Undo*** will undo the last operation.
- ***Legend*** will turn on below the graphics window a display of the correspondence between colours and filenames (automatically disabled if the selected files are 20 or more).

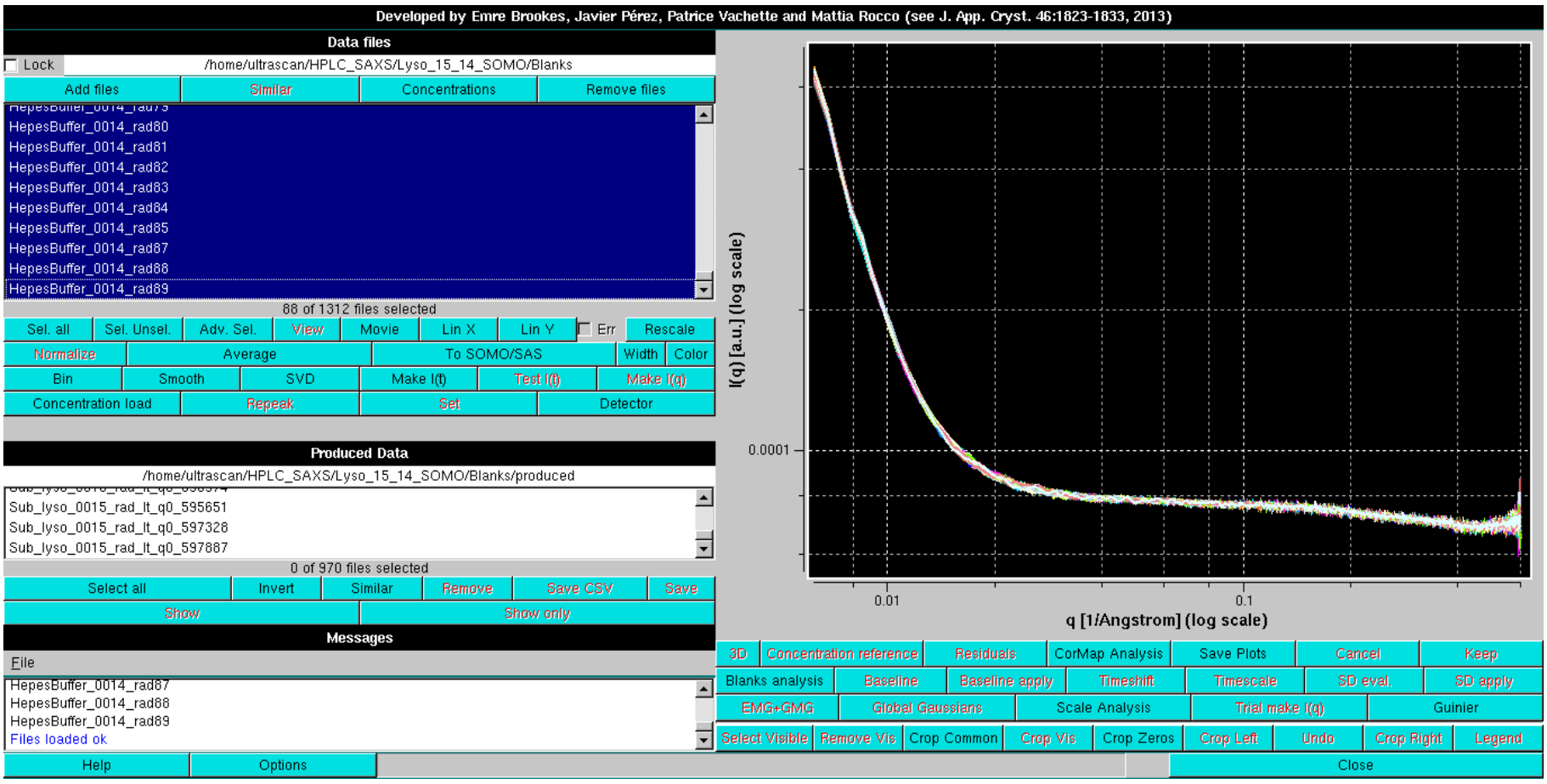
Of the top-row commands, two deserve already a comment at this point:

- ***CorMap Analysis*** will launch a pairwise similarity comparison between all currently selected datasets, according to our implementation of *P*-value computations and comparisons derived from the Correlation Map method developed by Franke *et al.* (Franke D, Jeffries CM, Svergun DI. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. Nature Methods, 12, 419-422, 2015). Specialized applications of the CorMap analysis are described at several points below. A general ***CorMap Analysis*** module description with examples can be found here.
- ***Save plots*** allows saving the data shown in any of the plots currently visualized in csv-formatted files. Pressing it will open a pop-up dialogue window where the location and the root filename for the cvs files can be set.

Visual inspection of the *I(t)* vs. *t* chromatograms can already hint at problems, such as in the lysozyme data here used as an example. In this case, it is evident that many *I(t)* vs. *t* chromatograms starting from the low-*q* region do not return to pre-peak elution baseline intensity values. Most likely, this is due to capillary fouling, and without proper correction these data would be mostly useless. For this, and for less evident cases, an Integral Baseline correction procedure has been devised

The **Integral Baseline** method is based upon the assumption that capillary fouling deposits are formed in proportion to the sample concentration while exposed to the beam, and that neither the buffer nor the instrumental background are contributing to this effect. That deposition on the capillary does occur is clearly proven by the fact that a steady SAXS signal is maintained even after completion of the protein elution. The theory underlying the Integral Baseline correction procedure can be found here.

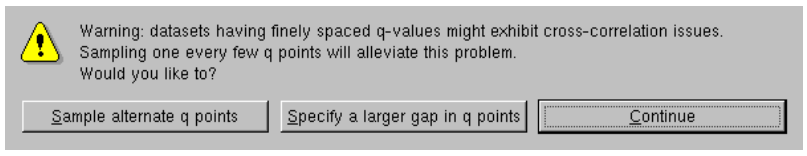
To help the user decide if a baseline correction is needed, and to find a proper region of SAXS steady state signal at the end of the chromatograms, the currently implemented Integral Baseline method requires an analysis on blank frames. These "Blanks" (*no less than **10 frames**, possibly at least **20 or more** must be available*) should have been collected well before the void volume, and should preferentially be the same ones that were then averaged and subtracted from all the data collected during the chromatogram development.



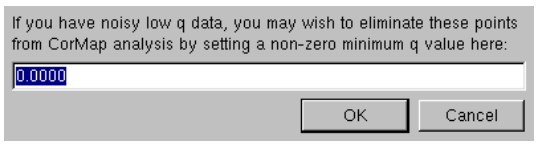
After Blanks files have been loaded using the ***Add files*** button (see above), their analysis is launched by pressing the ***Blanks analysis*** button. The module will automatically convert the *I(q)* vs. *q* frames into *I(t)* vs. *t* chromatograms:

The two vertical magenta lines and their corresponding fields at the bottom of the buttons' zone define the beginning and end regions for the Blanks analysis. By clicking on one of the fields and then moving the mouse on the grey-scale bar-wheel just below the graphics window, these limits can be changed. This can also be done in steps of a single frame by clicking on the "<" and ">" buttons placed at the extremities of the bar-wheel. Alternatively, the limits can be manually changed by entering a numerical value in their respective fields.

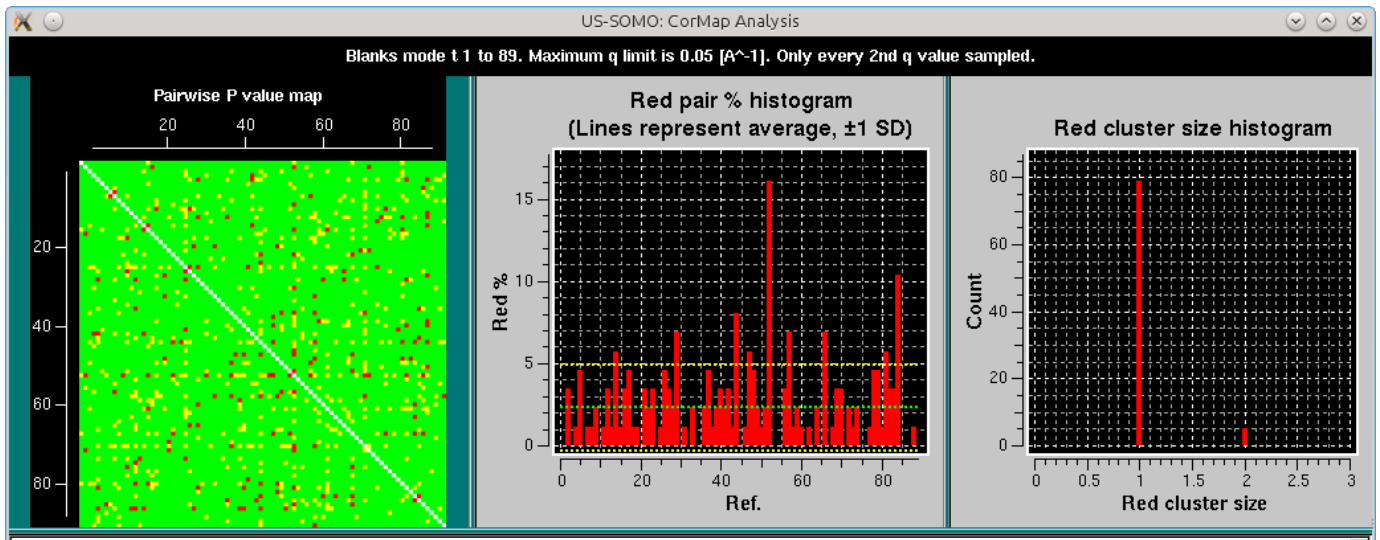
The Blanks analysis is performed by clicking on the ***CorMap Analysis*** button. This will launch a pairwise Correlation Map analysis (see here for a descrption of the CorMap implementation in the US-SOMO HPLC-SAXS module). Before the analysis is effectively launched, a pop-up panel will appear:



It was found during the implementation of the Blanks analysis that finely spaced q values might result in cross-correlation effects in the CorMap analysis (see also here). Therefore, this pop-up panel will allow to chose a sampling in q space to eliminate or at least alleviate this problem. Since usually a one-every-two values sampling is sufficent, this can be directly done by pressing the *Sample alternate q points* button. Larger sampling intervals can be chosen by entering an integer value after pressing the *Specify a larger gap in q points* button. If no sampling is wanted, the *Continue* button should be pressed. A second pop-up option will also allow to start the CorMap analysis above a chosen q_{min} value, to avoid including very noisy, low-q values in the analysis:



After these choices are made, the analysis is effectively launched, and the results are shown in a new pop-up panel (see here for a full description of the CorMap implementation):



The pop-up panel begings by reporting on the top bar the type of analysis (here "Blanks mode t 1 - 89"), the max q limit used (here 0.05 $Å^{-1}$), and the sampling used (here "Only every 2nd q value selected").

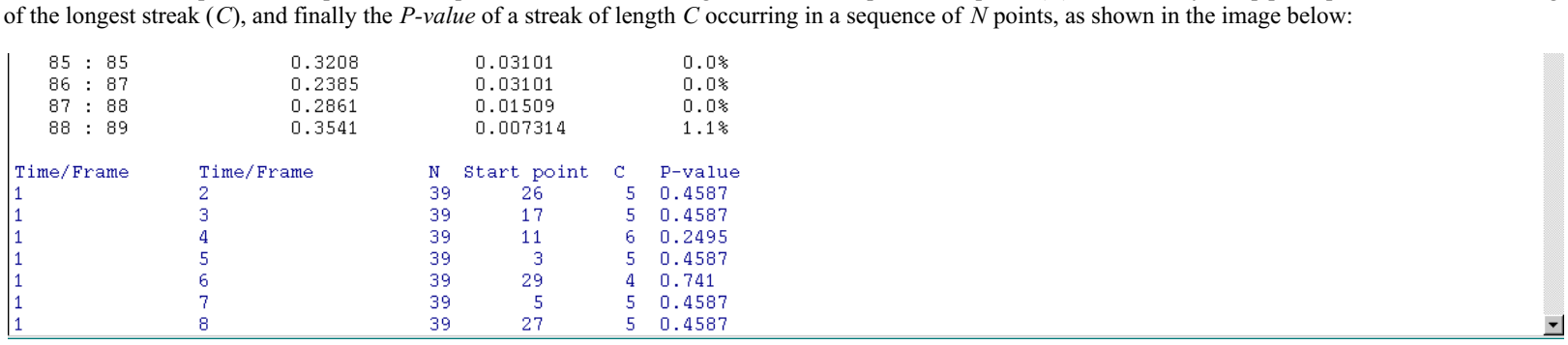
Three plots are present on the top of the panel:

- A **Pairwise P-value map**, a square matrix in which the square (i,j) contains the P-value resulting from the (i,j) curves represented by a three color code: green for $P \geq 0.05$, yellow for $0.01 \leq P \leq 0.05$, and red for $P < 0.01$. The labels on the axes correspond to the reference numbers given to the analyzed frames (see below).
- A **Red pair % histogram** plot as a function of the time/frame number, derived from the analysis of the pairwise P-value map in terms of the distribution of values between the three classes with an emphasis on the % of red squares.
- A **Red cluster size histogram**, derived from an evaluation of the average size of clusters of horizontally and/or vertically adjacent "red" squares. After careful examination of several indicators derived from the P-value map analysis, this was chosen as the most reliable to determine the global similarity between all frames of the considered subset.

The text area reports first the the pairwise P-map color definition, and then where to look for the correspondence between the axis ticks and the actual data. Then follows a summary of the most relevant data:

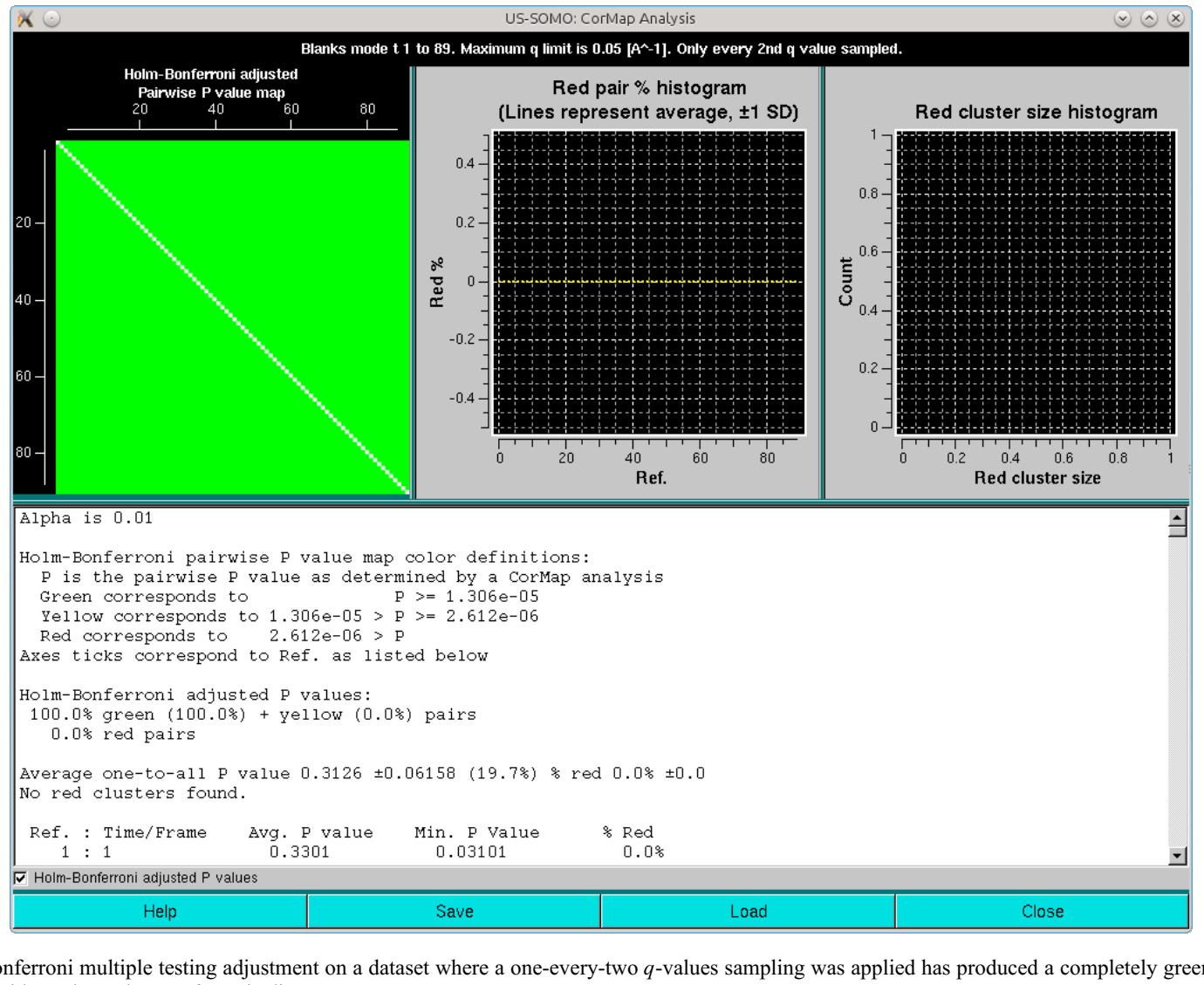
- A summary of the resulting P values pairs in terms of the green+yellow sum % and red % values.
- The average one-to-all P value ± its SD (in parentheses the % SD value) and the % of red values ± their SD (in parentheses the % SD value.)
- The red cluster count, their average size ± their SD (in parentheses the % SD value) and their average ± SD size as % of the total area.
- The red cluster maximum size (with its % in respect to the total area), the number of occurences of that cluster size, and the position of the first occurence.

Below these summaries, the first list reports the correspondence between the "Reference numbers" (**Ref**) assigned to each dataset (here frames) and its "real" name. This was introduced to avoid having to deal with complicated names in the axes legends on the plots (in this case, the frame numbers that were extracted from the *I(t)* vs. *t* filenames start at "1", so they are equal to the "Ref" numbers). The list also reports for each frame the *Avg. P value*, the *Min. P value*, and the *% Red* points.
At the end of the first list, a second list reports all the pairwise comparisons results, including the number of points compared (*N*), followed by the q point position where the longest streak occurs (*Start point*), then the length of the longest streak (*C*), and finally the *P-value* of a streak of length *C* occurring in a sequence of *N* points, as shown in the image below:

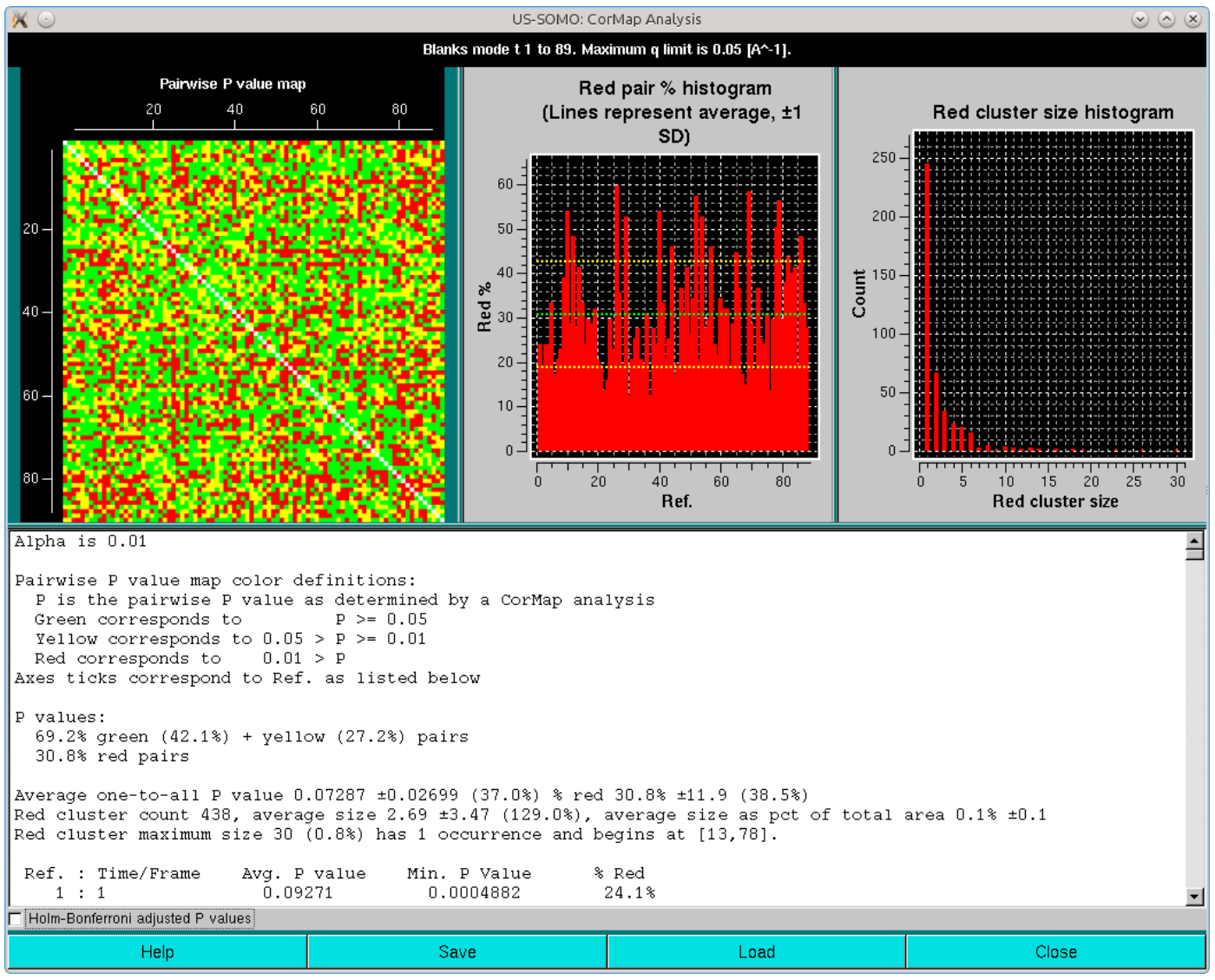


85 : 85		0.3208	0.03101		0.0%
86 : 87		0.2385	0.03101		0.0%
87 : 88		0.2861	0.01509		0.0%
88 : 89		0.3541	0.007314		1.1%
Time/Frame	Time/Frame	N	Start point	C	P-value
1	2	39	26	5	0.4587
1	3	39	17	5	0.4587
1	4	39	11	6	0.2495
1	5	39	3	5	0.4587
1	6	39	29	4	0.741
1	7	39	5	5	0.4587
1	8	39	27	5	0.4587

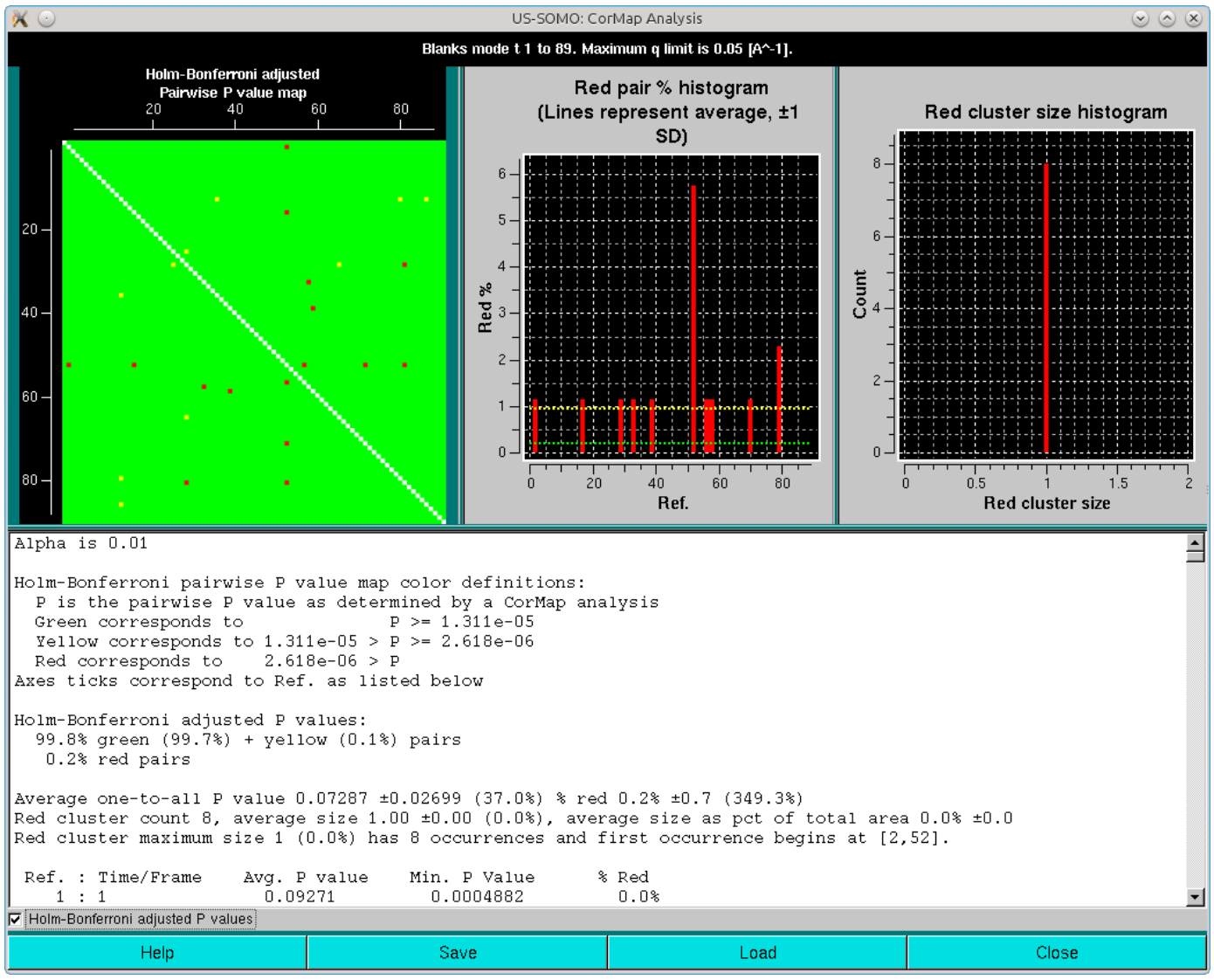
At the end of the text area, a checkbox is present. If selected, the paiwise analyses will be adjusted for multiple testing using the Holm-Bonferroni approach (Holm, S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6:65-70, 1979; see here):



Note that in this case the Holm-Bonferroni multiple testing adjustment on a dataset where a one-every-two *q*-values sampling was applied has produced a completely green pairwise *P*-values map. Without sampling, this is what is obtained without the Holm-Bonferroni adjustment:



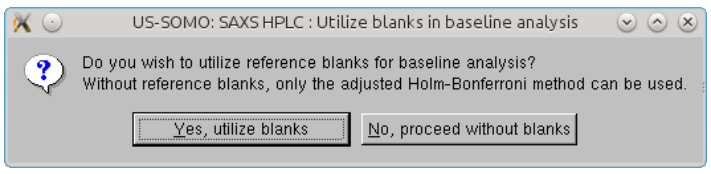
and with Holm-Bonferroni adjustment:



All data listed in the CorMap analyis pop-up window can be saved in a csv-type file with the ***Save*** button. Previously analyzed datasets can be recalled with the ***Load*** button.

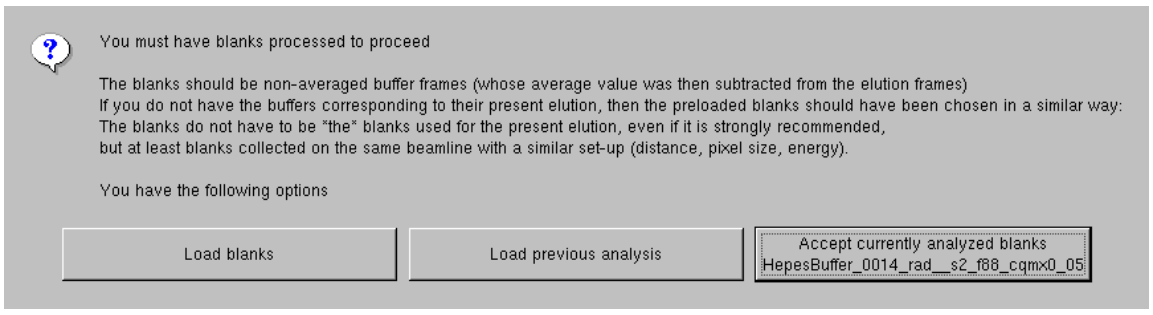
After closing the CorMap analysis window, the Blanks data can be accepted by pressing the ***Keep*** button. ***Cancel*** will instead discard the current CorMap analysis.

The Integral Baseline analysis of the actual sample frames can then begin. Contrary to what was required in our previously developed Integral Baseline method, the current version requires that **all** *I(t)* vs. *t* chromatograms must be selected before hitting the ***Baseline*** button.
In any case, on pressing ***Baseline*** a first pop-up warning message will always appear:



This allows the user to proceed without using a Blanks reference set to judge if a stable baseline has been reached at the end of the chromatograms. In this case, only the Holm-Bonferroni adjusted pairwise comparison will be used.

If Blanks are to be used, a second pop-up warning message will appear:



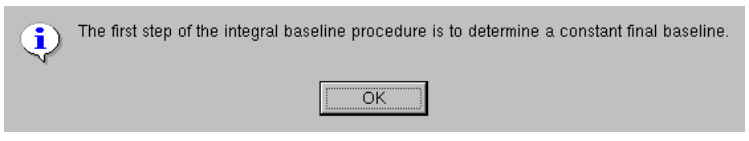
alerting that a blanks analysis is needed to proceed any further, and offering up to three options:

- *Load blanks*, which will allow to load *I*(*q*) vs. *q* Blanks files to be subjected to the analysis as described above;
- *Load previous analysis*, which will allow to select a previously saved Blanks analysis cvs file;

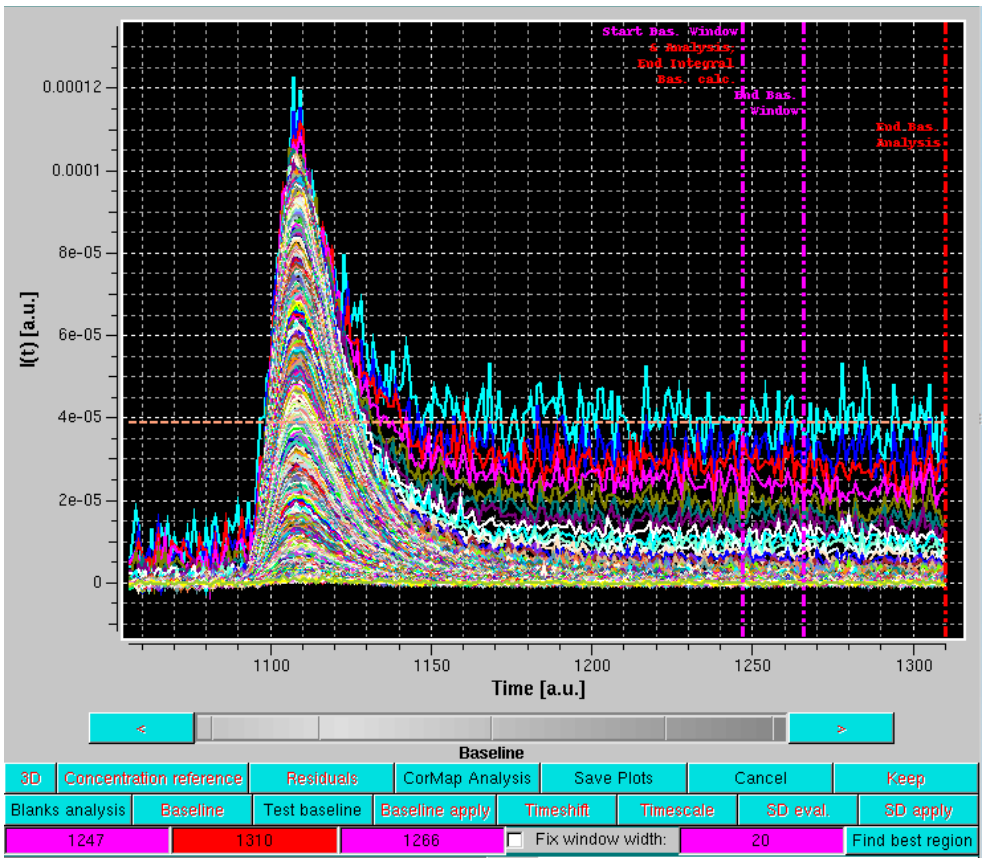
and, if Blanks were analyzed during the current session,

- *Accept currently analyzed blanks* (followed by the filename of the currently analyzed blanks).

Another pop-up will then appear, reminding that the first step in the Integral Baseline procedure is to find a final region of constant intensity:



After pressing *OK*, the graphics window will present all the selected *I(t)* vs. *t* chromatograms and switch to the *Baseline* mode of analysis:



As shown in the image above, this superimposes to the selected chromatograms three vertical lines on the right side, the last two lines of buttons under the graphics window are replaced by three colored fields (magenta-red-magenta), and a dashed line is drawn horizontally (orange). In addition, a *Fix window width* checkbox with its associated magenta-colored field in now present (**default: 20 frames, unchecked**), as well as a new ***Find best region*** button.
The first vertical magenta line, which by default is positioned at 75% of the available frames, has multiple usages:

- It defines the starting frame of a sliding window over which a **CorMap** analysis will be carried out;
- It will also define the first frame where the **CorMap** analysis will start;
- When the Integral Baseline is calculated, it will be the end point of the computations (but the calculated baseline, which will be constant beyond this point, will be then subtracted up to the last frame).

The second vertical magenta line defines the end of the sliding window (**default: 20 frames** beyond the first magenta line).

The vertical red line defines the end for the sliding window analysis (**default position: 5 frames** from the end of the available frames).

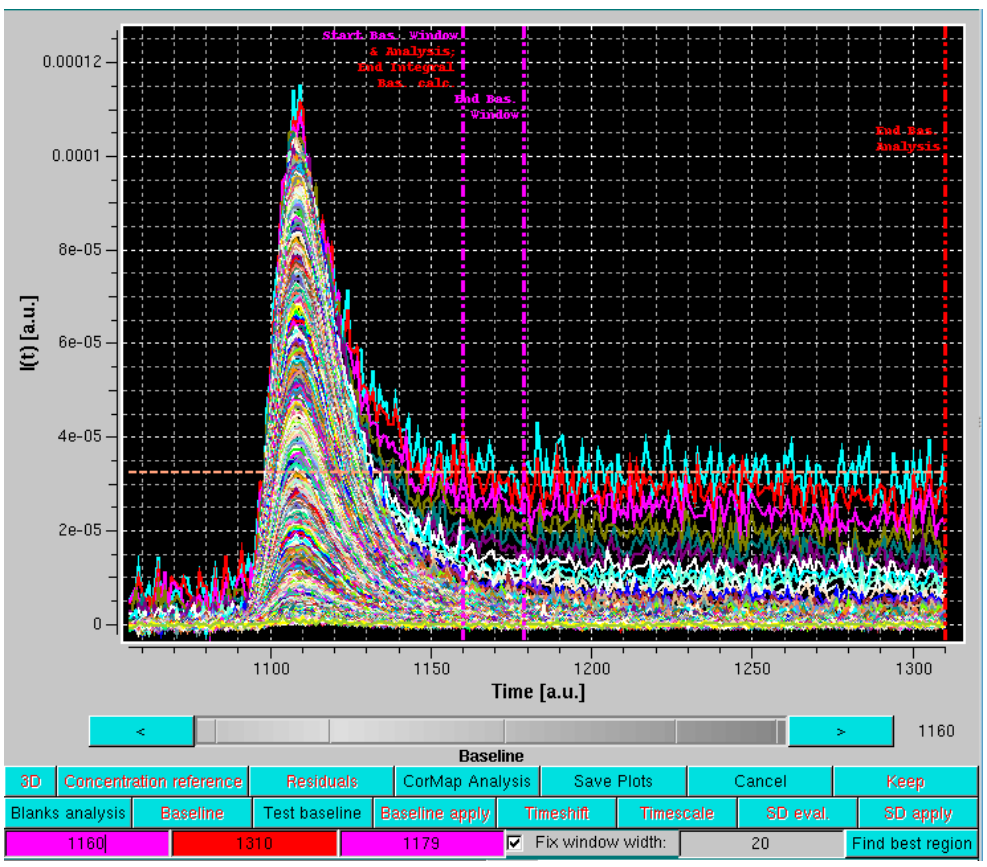
The horizontal orange line represents the average intensity across the current window of the lowest *q*-value among the selected *I(t)* vs. *t* chromatograms.

It is important to remind that the baseline is set to be at zero at the beginning of the data on the left side.

The positions of the three vertical lines are indicated in the three background color-coded fields. By cliking on one of the fields, the corresponding vertical line position can be changed using either the grey-shades bar-wheel, or the "<" and ">" buttons at its sides. Manual values can be also entered.

If the *Fix window width* checkbox is not selected, moving either of the two vertical magenta lines will also change the width of the sliding window.

It is then best to first define a window width by moving either one of the vertical magenta lines, and then fix it by selecting the *Fix window width* checkbox. At this point, the entire window can by positioned by using either of the two vertical magenta lines. It is suggested to position it in a region where there is still some visible intensity decay, as shown below:

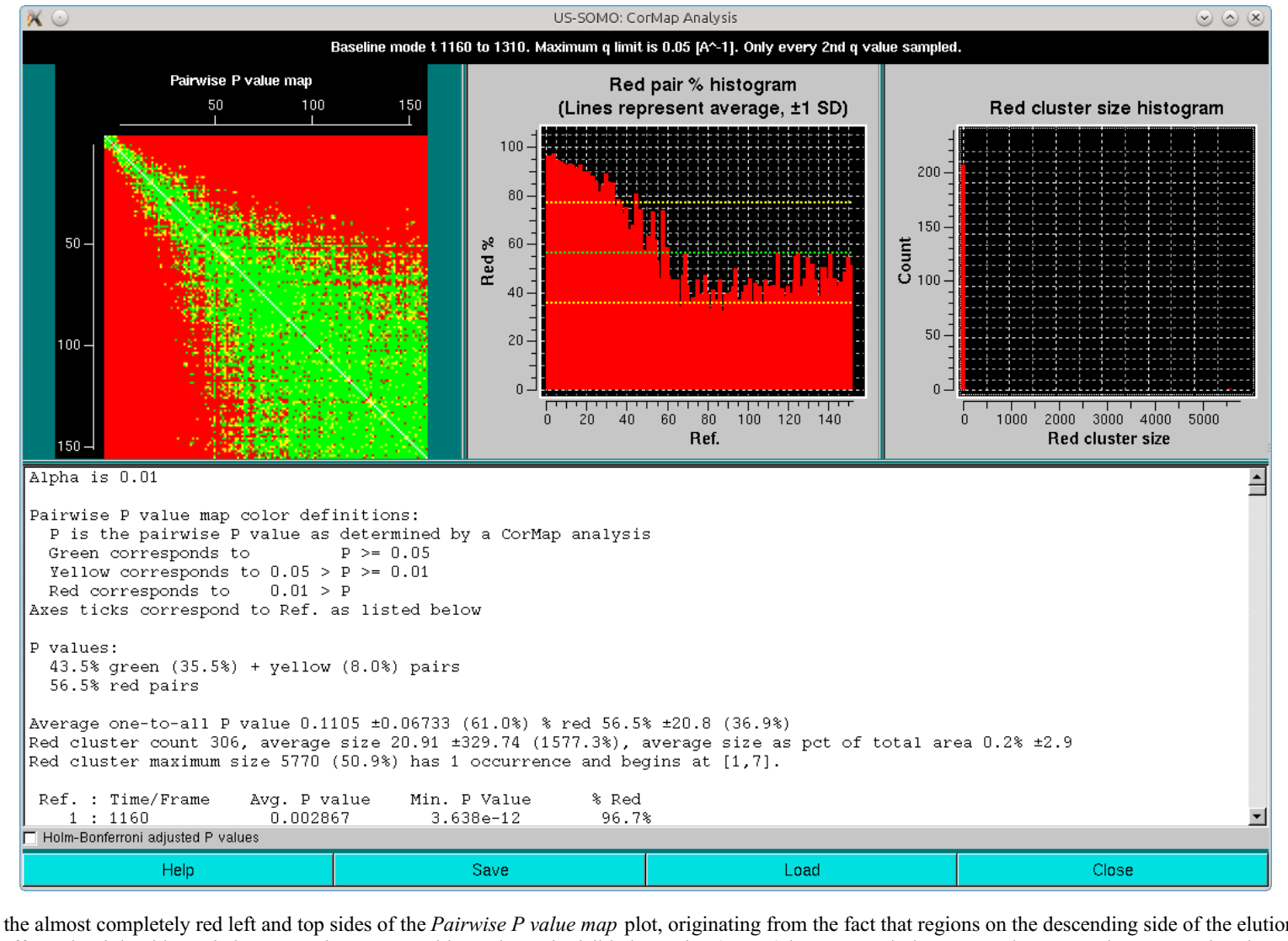


The baseline analysis is then completed by pressing the ***Find best region*** button.

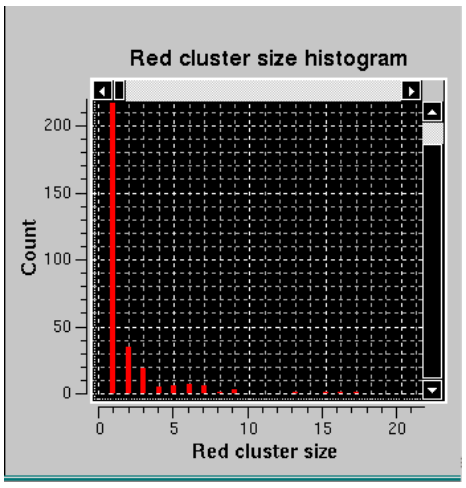
This will launch a special **CorMap** analysis in which first a global CorMap calculation will be carried out between the entire range of frames from the first vertical magenta line to the vertical red line. Subsets of this CorMap analysis corresponding to the sliding window regions will then be extracted and compared with the average of all possible CorMap analysis results extracted from the pre-analyzed Blanks data for a sliding window of the same size.

In addition, the analysis will calculate the integrated average intensity at each frame of all the $I(q)$ values from the minimum q-value selected up to the q_{max} defined in the **Options** panel (**default: 0.05 Å^{-1}**).

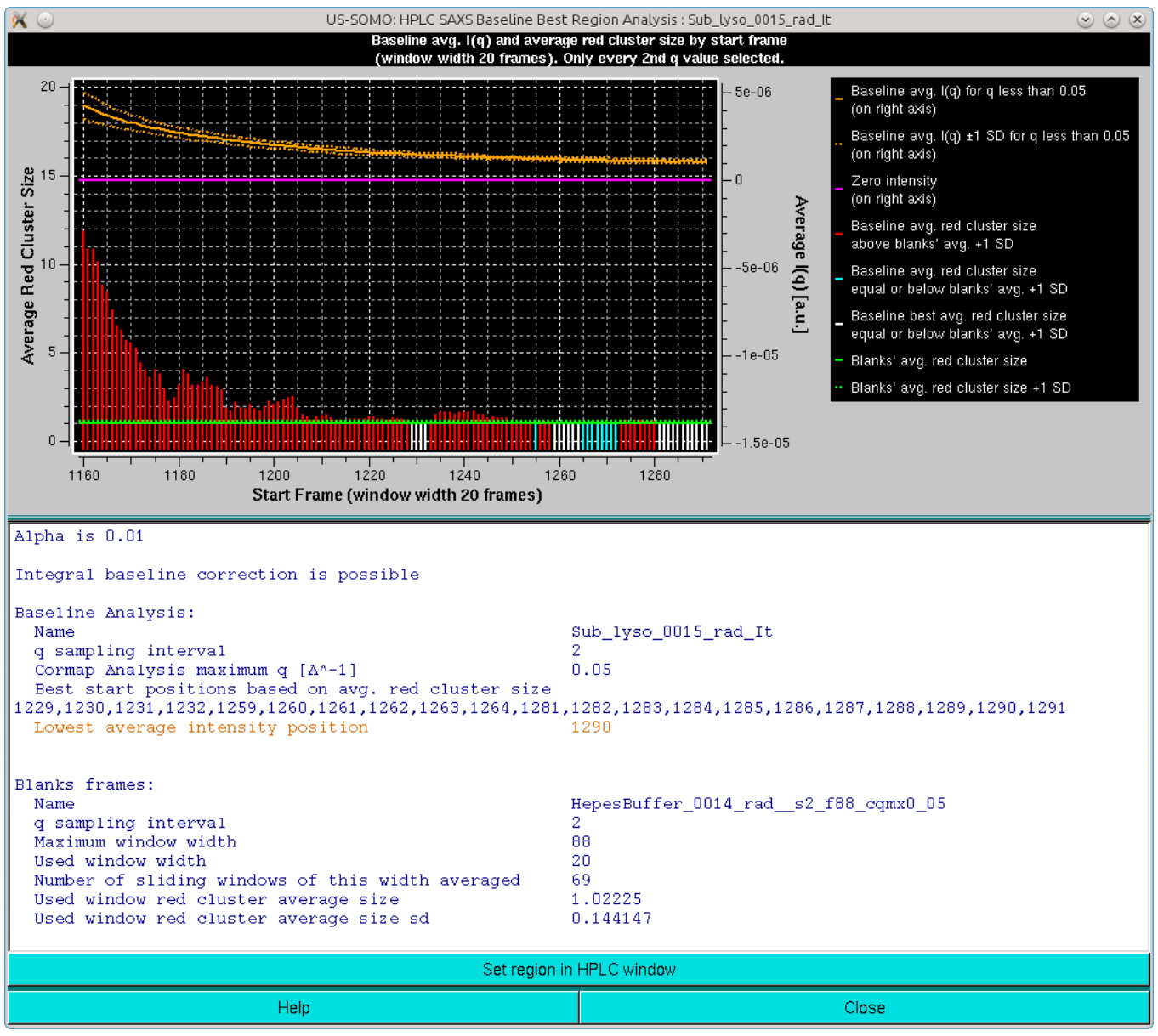
The results will appear in two pop-up panels. The first one is analogous to the one appearing after the Blanks analysis:



Here, it can be appreciated the almost completely red left and top sides of the *Pairwise P value map* plot, originating from the fact that regions on the descending side of the elution peak were included in the analysis. This also heavily affects the right-side *Red cluster size histogram*, with an almost invisible huge size (≈5000) but extremely low count cluster greatly compressing the scale. If we zoom on the low red cluster size region, this is what becomes visible:



But most relevant is the second pop-up panel that will appear on top of the first:



The graph in this panel is composed of two plots, both as a function of the starting window position. The bottom histogram (left-side *y*-axis scale) reports the average red cluster size for each window in the

sliding window ensemble. The horizontal green solid line defines the Blanks average red cluster size for all possible windows of the same size as the sliding window utilized for the Sample analysis (the dotted line represents + 1 SD). The bars in the histograms are colored red when they are above the Blanks + 1 SD value, while cyan and white when they are ≤ the Blanks + 1 SD value, with the white being the lowest value(s) (equal values are possible).

The top plot (right-side *y*-axis scale) reports the averaged $I(q)$ for $q \leq$ the q_{max} value (0.05 $Å^{-1}$ by default, as set in the **Options**), as the solid orange line, with the dotted orange lines representing ±1 SD. The solid magenta line defines the zero value expected for blanks-subtracted data when only buffer is present.

The goal of this combined analysis is twofold:

- To find region(s) of steady-state intensity;
- To compare the average integrated $I(q)$ intensity with the zero value expected for pure buffer conditions.

The first message appearing in the text region concerns the second of the points listed above.

- If the average integrated $I(q)$ ±1 SD vs. starting frame position plot is **always above** the zero reference line **and** there is at least one occurrence of an average red cluster size being **equal or less** than the Blanks' average red cluster size +1 SD, the message "*Integral baseline correction is possible*" appears.
- If the average integrated $I(q)$ ±1 SD vs. starting frame position plot is **always above** the zero reference line **but** the average red cluster sizes are always **higher** than the Blanks' average red cluster size +1 SD, the message "*Integral baseline correction would be possible, but no steady-state end region has been identified*" appears.
- If the average integrated $I(q)$ ±1 SD vs. starting frame position plot **crosses and goes below** the zero reference line, **irrespective** of the average red cluster size status, the message "*Integral baseline correction is not recommended*" appears.

Below the summary sentence, a first report of the baseline analysis is printed. It contains:

- The root name of the files analyzed;
- The q sampling interval;
- The CorMap analysis q_{max};
- A list of the best starting positions for end frames (that span a length equal to the chosen sliding window size) where the average red cluster size is within 1 SD similar to those of the Blanks (white bars in the histogram; if no white bars are present, green bars will be listed; if neither white of green bars are present, a single, lowest count bar will be listed, colored yellow in the histogram);
- In orange color, the frame position of the lowest average integrated intensity.

A second block of information is then printed further down, containing Blanks-related information:

- The root name of the Blanks frames files utilized in this analysis;
- Their q sampling interval;
- The maximum available sliding window width;
- The used sliding window width;
- The number of sliding windows averaged;
- The red cluster average size for the used sliding window;
- And its associated SD.

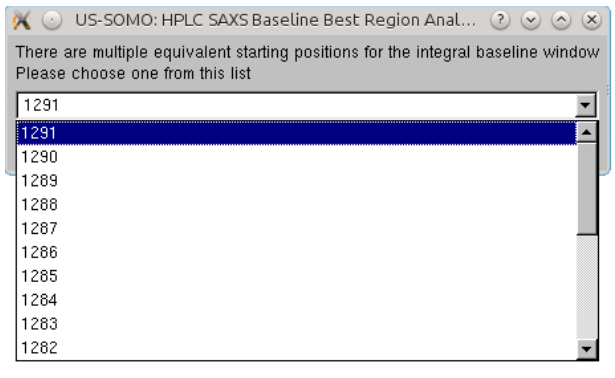
The user should then examine the information provided and consider the suggestion made at the top of the text window:

If the averaged integrated intensity -1 SD reaches at some point the zero line, then no Baseline correction is likely necessary.

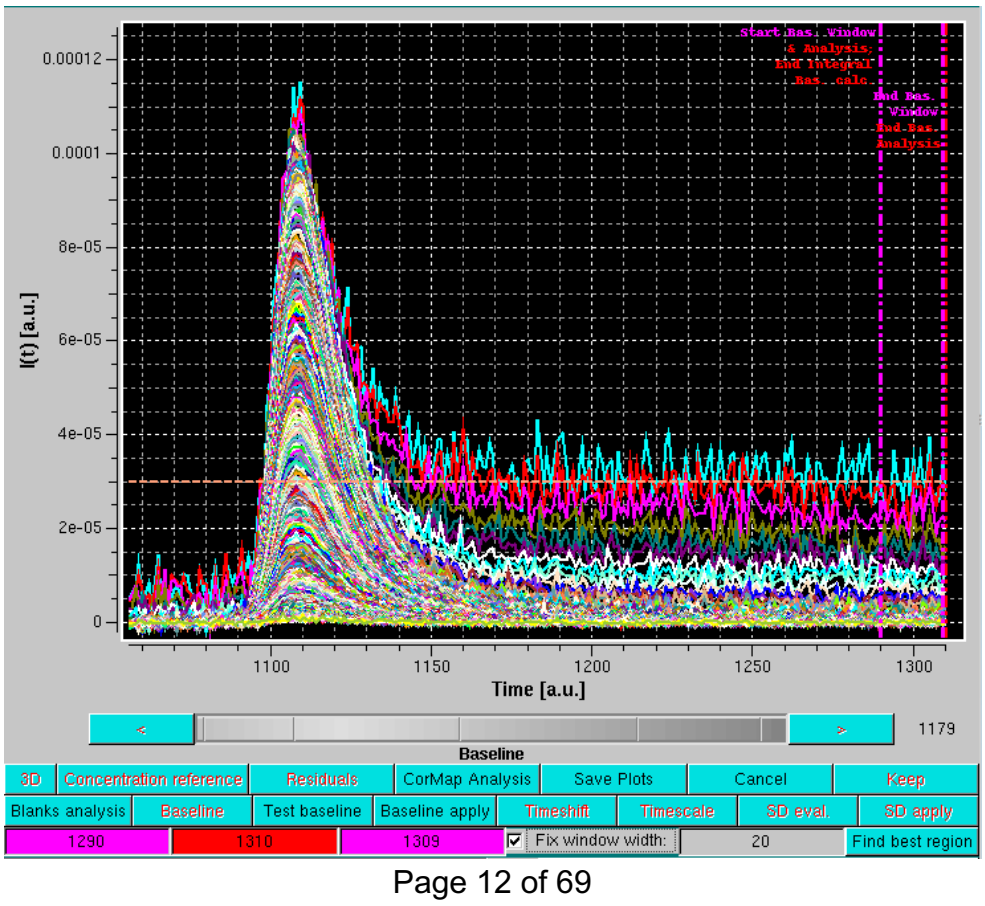
If the averaged integrated intensity -1 SD at some point crosses and goes +1 SD below the zero line, then other issues might be present, such as incorrect Blanks subtraction, or drifting problems. In the latter case, a Linear Baseline correction might be indicated (see here).

If the averaged integrated intensity is always above +1 SD of the zero line, then an Integral Baseline correction could be necessary. The second condition warranting it is that there is an end region where a sufficient number of Sample frames (equal to the sliding window size by definition) is judged by the average red cluster size of the Pairwise P value Map to be similar within +1 SD of the average Blanks frames. Those are the starting frames listed in the first block of summary information, but beware of the presence of a single yellow-colored starting frame: it means that no frames passed the stringent Pairwise P value Map average red cluster size test, and that the listed frame is only the one having the lowest average red cluster size (which can be much higher than the Blanks average red cluster size). The user could try repeating the analysis using a different (smaller) sliding window size. Check also the starting/ending positions to be sure to include an appropriate end region for this analysis.

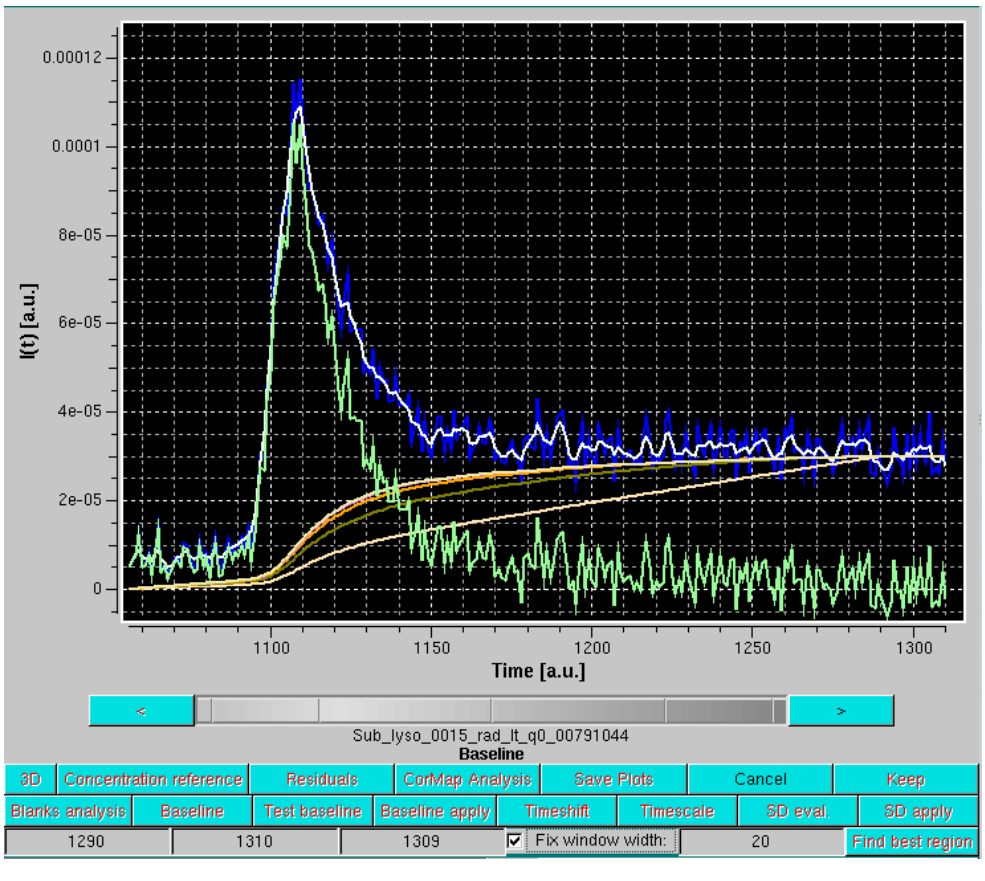
When both the Integral Baseline applicability conditions are met, the user could automatically transfer the position of the window into the Baseline module, by clicking on the ***Set region in the HPLC window*** bar at the bottom of the an analysis window. This will open a pop-up panel:



listing all possible starting positions for the baseline window, beginning with the farthest one. The user should pick a position matching with the lowest average integrated intensity frame position (orange colored text). If more than a lowest average integrated intensity frame is available, it is advisable to pick an earlier one, to avoid a potential undercorrection when integral baseline subtraction is performed. Once a position is chosen, clicking on *OK* will transfer it to the Baseline module panel:



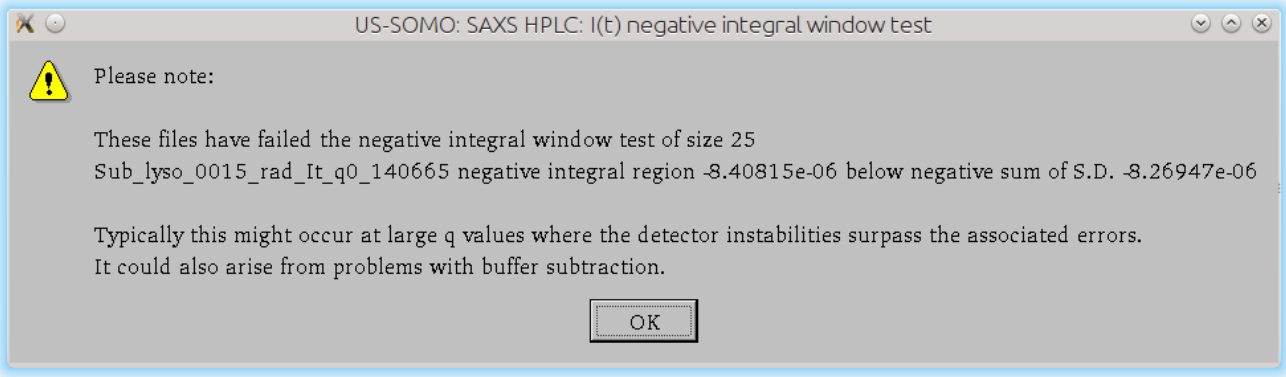
To verify what the Integral Baseline will effectively produce, ***Test Baseline*** should be launched. This will show in scroll mode every original *I(t)* vs. *t* chromatogram, a smoothed version using a Gaussian smoothing kernel of 2*n*+1 points (where *n* is set in the Options panel, with *n* = 3 as default), the iterations in the Integral Baseline computation (whose number is also set in the Options panel, with 5 as default), and the final, integral baseline-corrected chromatogram:



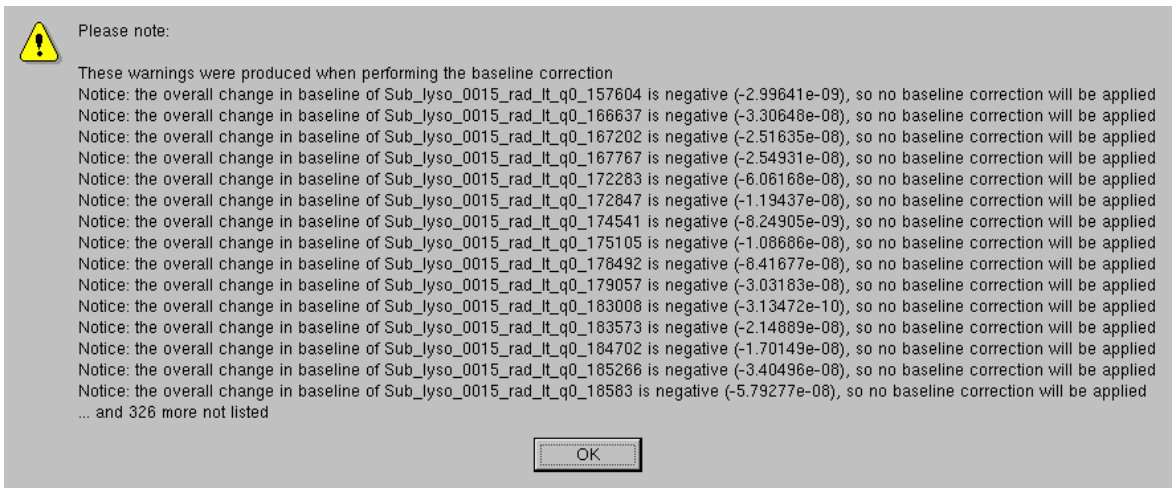
In the example shown above, blue is an original *I(t)* vs. *t* chromatogram at *q* = 0.00791 Å^{-1}, white its smoothed version with the default settings, cream, olive green, orange and pale yellow are integral baseline iterations 1, 2, 3 and 5, respectively (the 4th iteration is not visible, completely superimposed by the 5th), and light green is the final, integral baseline-corrected *I(t)* vs. *t* chromatogram. The Gaussian smoothing is applied to remove large oscillations in the original *I(t)* vs. *t* chromatogram, giving rise occasionally to values below the current integral baseline iteration, leading then to addition rather than subtraction in the computations. The final integral baseline is then subtracted from the original *I(t)* vs. *t* chromatogram, not the smoothed one.

As can be seen in the example above, the procedure appears to have produced a reasonable correction. All *I(t)* vs. *t* chromatograms can be checked in the ***Test baseline*** mode, which can be abandoned by pressing ***Cancel***.

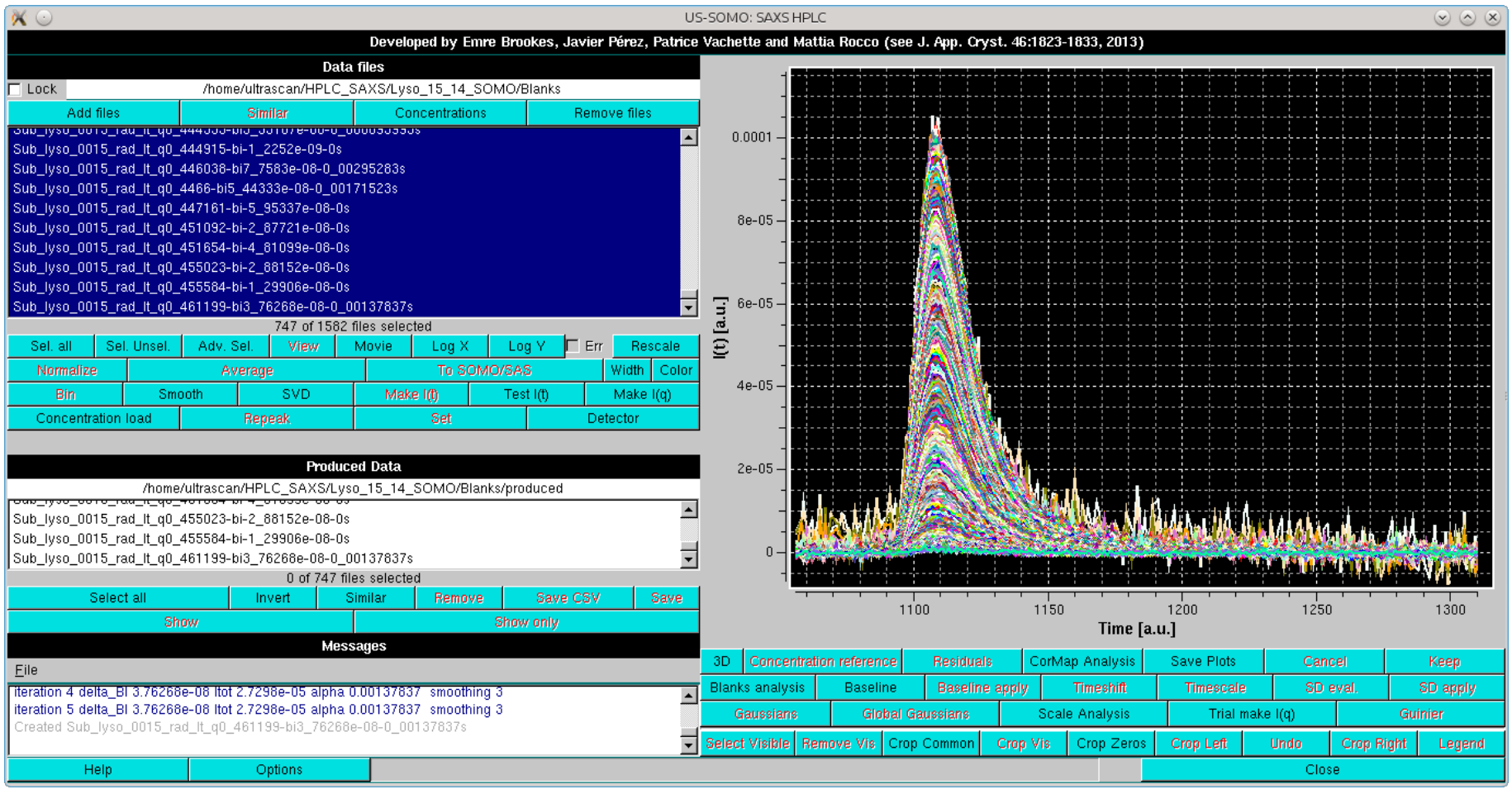
The Integral Baseline procedure can now be applied to all selected *I(t)* vs. *t* chromatograms by pressing ***Baseline apply***. If files that failed the negative regions test within a sliding window (of 25 frames in this case) where the sum of the intensity is less than the negative of the sum of the corresponding SD values over the window are present, this message will again appear:



After pressing *Ok*, all the integral baselines will be computed and then subtracted from the *I(t)* vs. *t* chromatograms. When the overall computed change in baseline is found to be negative, no baseline correction is applied. A pop-up panel will alert the user listing the first 20 such occurences and giving the number of all the others found:

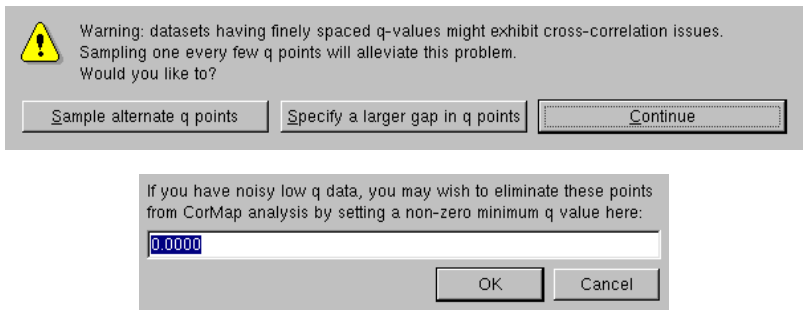


Each resulting baseline-subtracted chromatogram will have a "-bi" added after the *q* value and an "-s" at the end of the filename to indicate that an integral baseline subtraction was applied (if a linear baseline option is used, the first label will be "-bl"). The numerical value of the overall change in baseline and the alpha value (for an explanation alpha see here) are also added to the filename of the produced files, as shown in the ***Data files*** panel. Files where no baseline was subtracted will have a "0s" at the end of the filename:

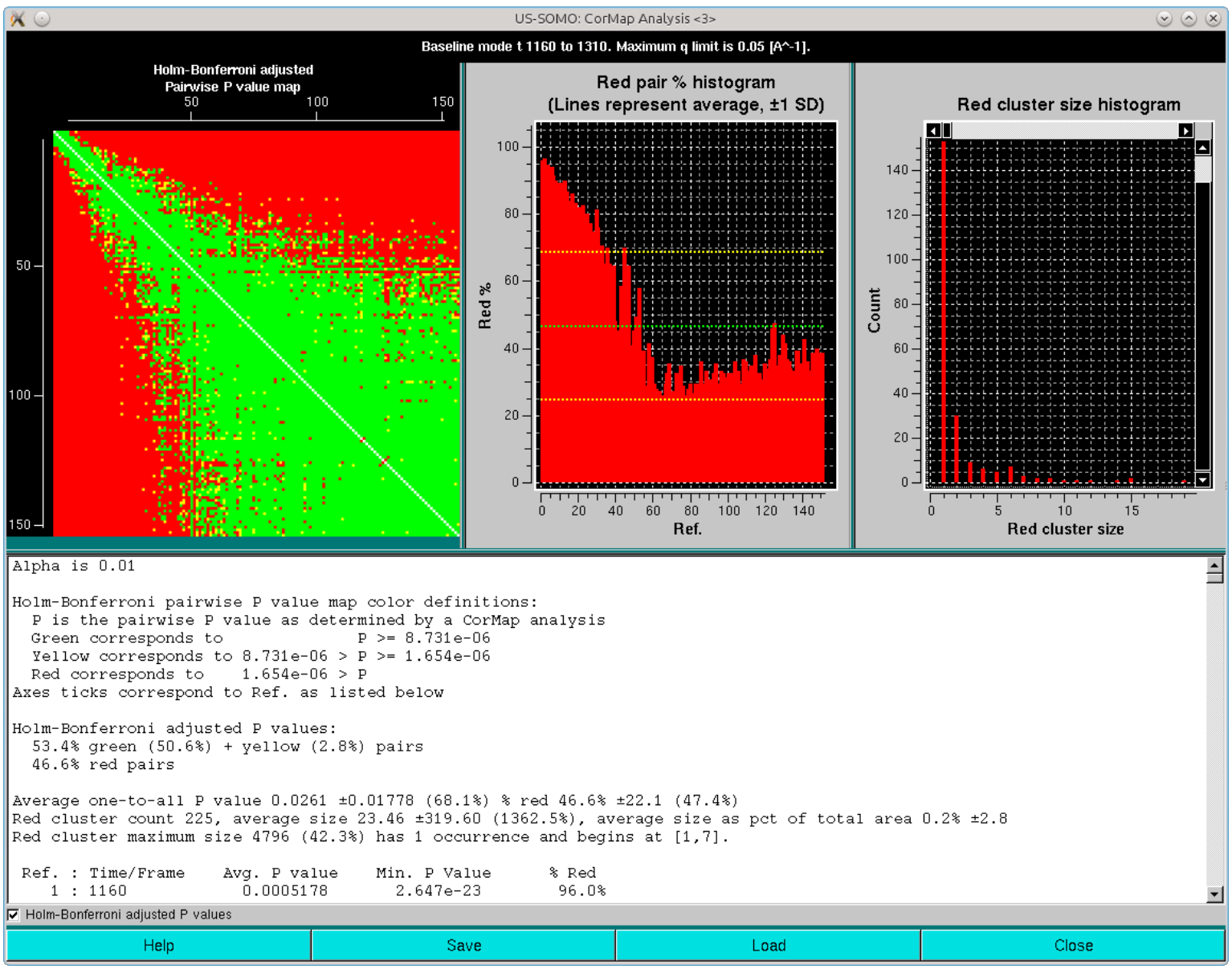


If the baseline procedure without Blanks was selected, the Baseline module will proceed slightly differently.

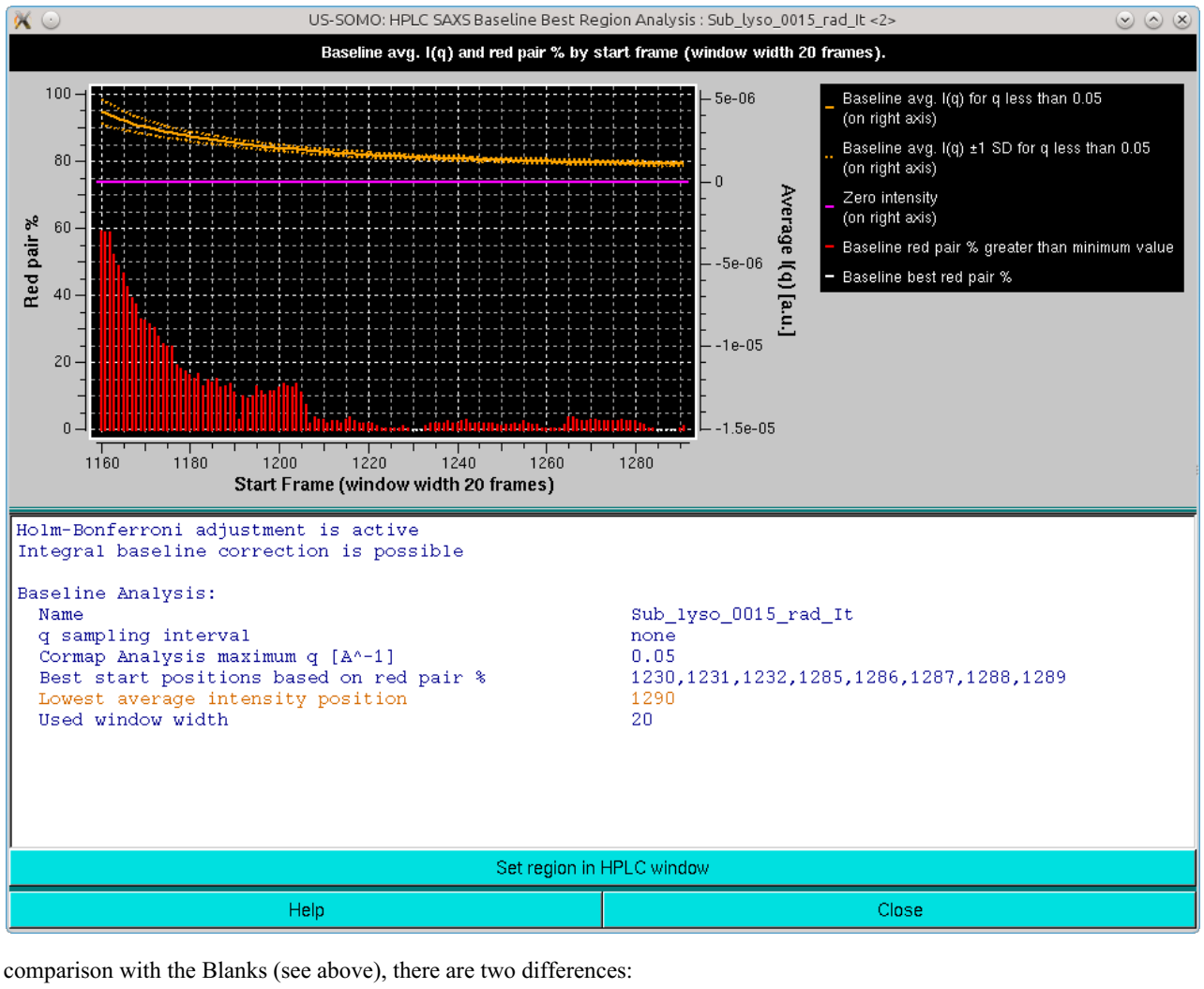
First, the "sampling" pop-ups will appear, since in this case the sampling is not set by what was used for the Blanks:



In the following example, no sampling was applied. The sliding window size and the beginning-end of the baseline analysis region are set as with the procedure with Blanks (see above), and the ***Find best region*** button is then pressed. The CorMap results will this time be displayed with the Holm-Bonferroni checkbox automatically selected:



The second pop-up will also appear:

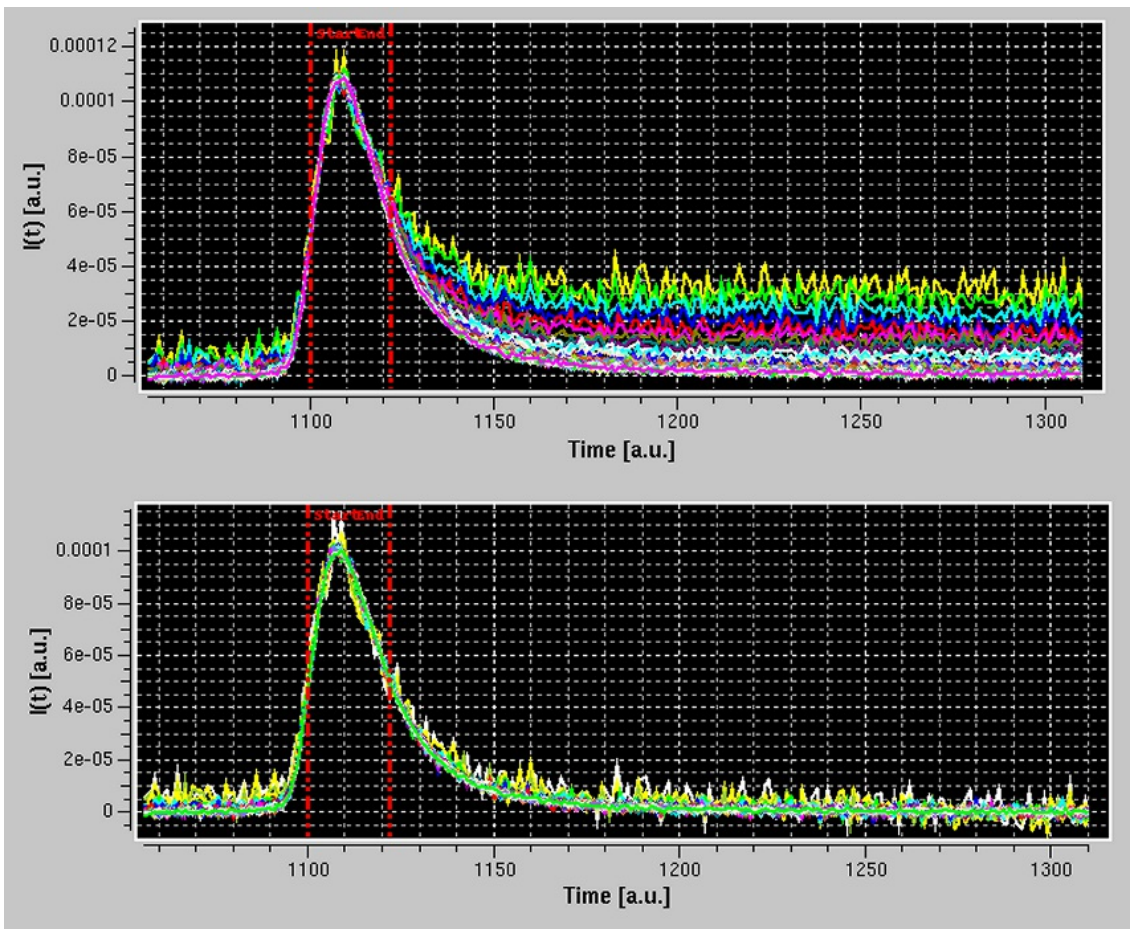


In respect with the analysis including a comparison with the Blanks (see above), there are two differences:

- The histogram reports the Red pair % instead of the Average Red Cluster size. This is because with the Holm-Bonferroni adjustment the formation of red cluster is severely inhibited.
- The Blanks average line is obviously absent. The cut-off to mark the histogram bars white is set to 1%. If no window reaches this level, the lowest window will be marked with a yellow color in the histogram.

All other functionalities are as with the analysis including the Blanks comparison.

A first assessment of the results of the Integral Baseline subtraction can be done by comparing two identical subsets of $I(t)$ vs. t chromatograms before and after correction. In the figure below, the subsets are from q = 0.00791 to q = 0.05029 Å^{-1} scaled on each other in a frame interval corresponding to the half-height of the peak, before (top panel) and after (bottom panel) baseline subtraction:

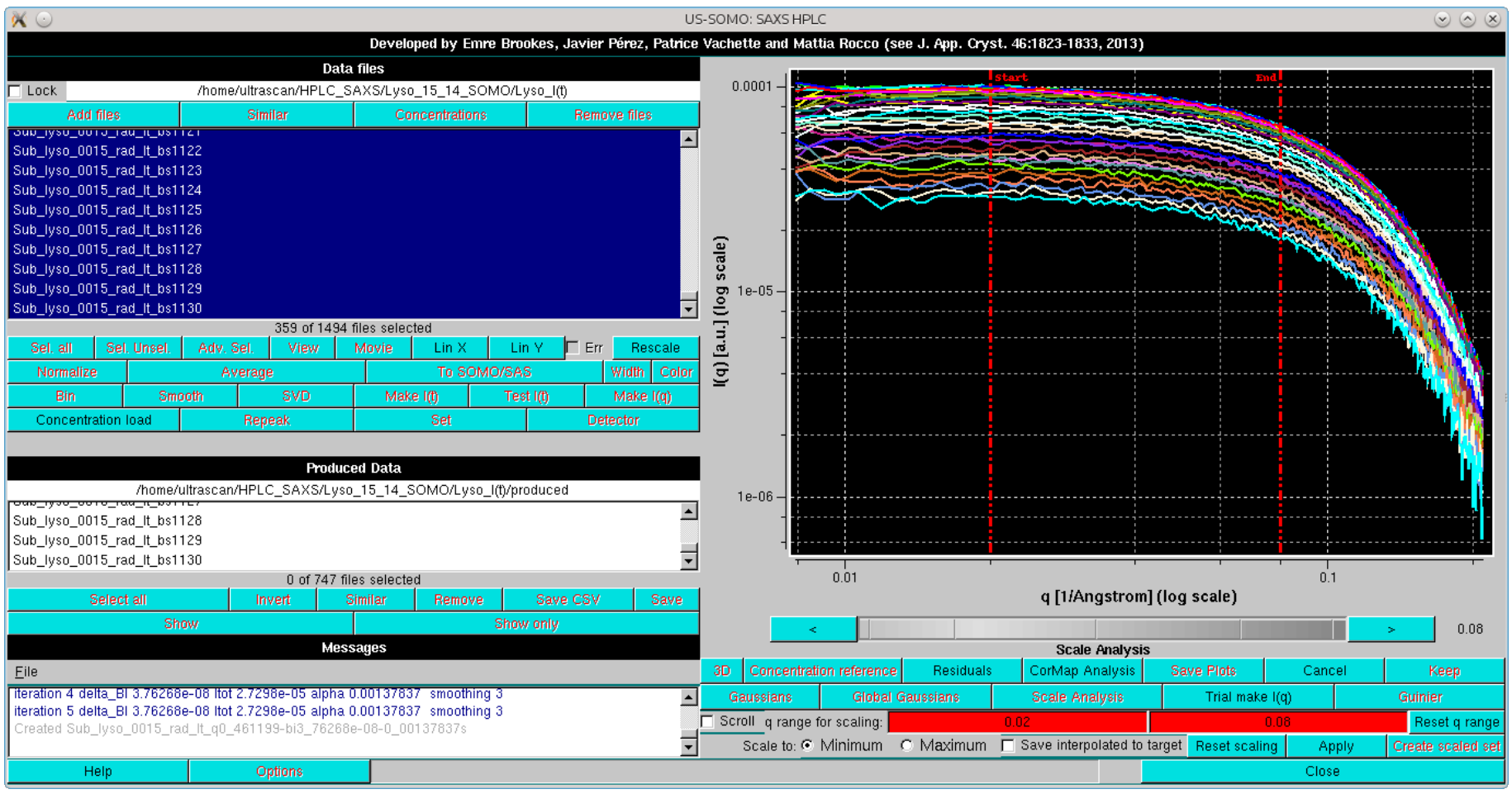


The fact that the right-hand sides of the peaks are nicely superimposed after baseline subtraction validates a posteriori the procedure used to build the baseline, since for a single species the elution peaks at different q-values should be strictly proportional to each other.

More checks of the Integral Baseline subtraction correctness can be performed using the ***Trial make I(q)*** mode. In this mode, the selected $I(t)$ vs. t chromatograms are temporarily transposed back into $I(q)$ vs. q frames and can be analyzed by either scaling or Guinier approximation utilities. In the example below, we have selected a subset of q-values from ≈0.008 to ≈0.21 Å^{-1}, and we have pressed the ***Trial make I(q)*** button:

This will bring up again the gray-shades wheel-bar and change the two lowermost bars with the buttons below the graphics window. At the bottom, a *Time range for I(q):* label will appear, followed by two fields with red background indicating the region subjected to the ***Test I(q)*** procedure. The limits can be changed by either clicking on each red-colored field and then using the gray-shades bar-wheel at the top, or on the "<" and ">" buttons placed at its sides. Alternatively, if a region was pre-selected with the mouse, it can be applied by clicking on the ***Vis. range*** button. In this example, we have set the *Time range for I(q)* limits from frame 1100 to 1130.

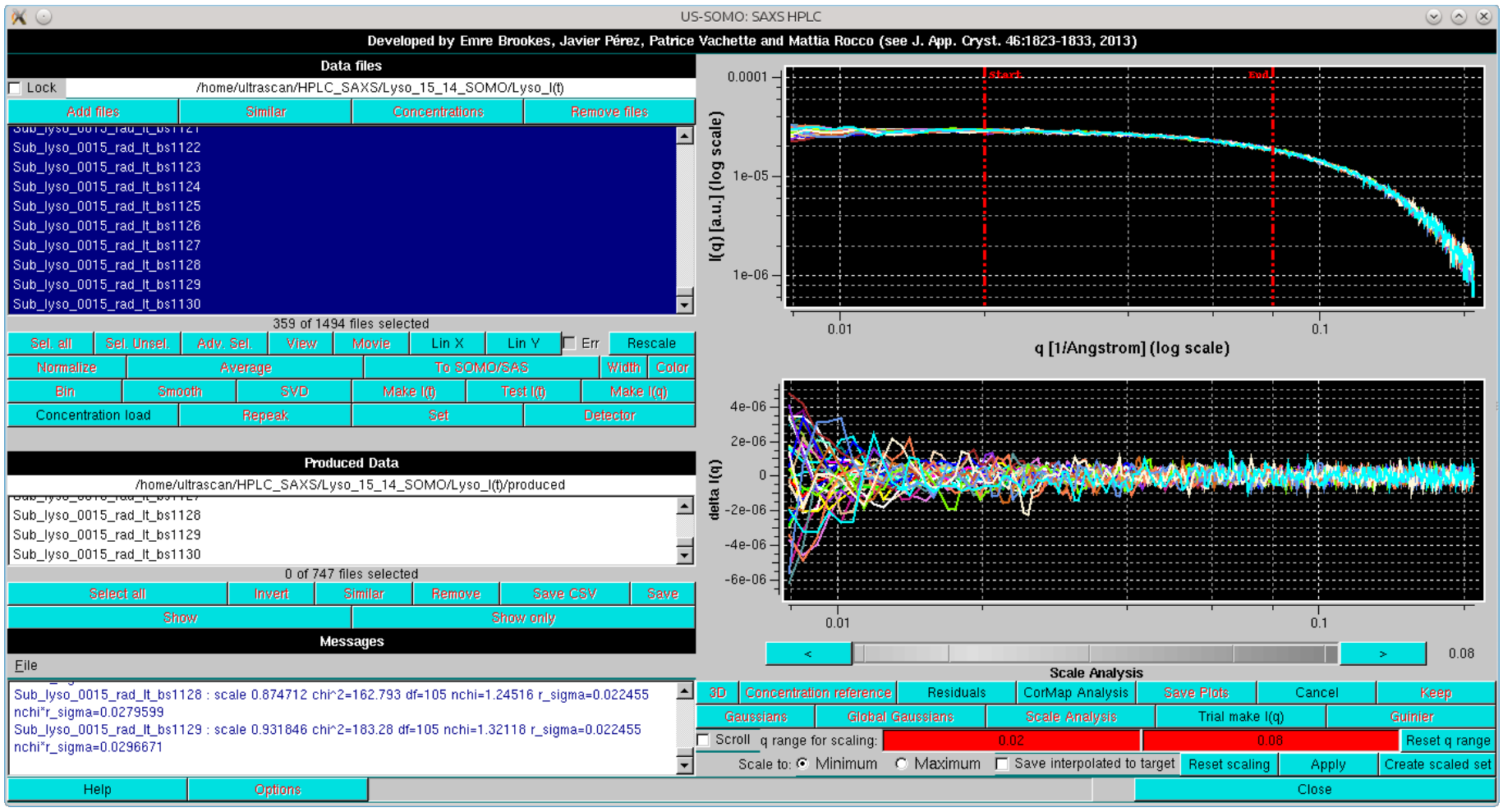
Two operations can be then performed. Pressing the ***Scale Analysis*** button on the row above will change the layout in this way (note that, by pressing ***Log X*** and ***Log Y*** on the left-side commands panel, both axes have also been changed to $\log_{10}$):



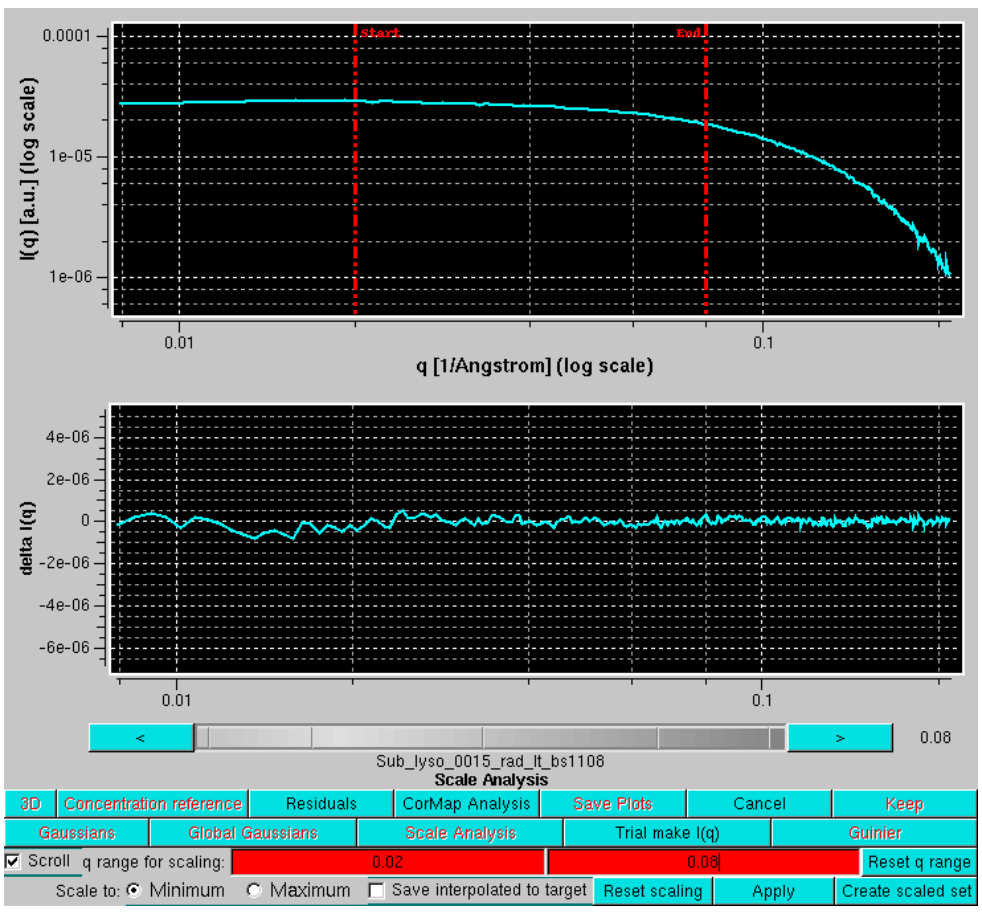
The two lowermost rows now display the tools for scaling the back-generated temporary $I(q)$ vs. q curves on top of each other. The two red-background fields now indicate the actual *q range for scaling*, which can be adjusted by clicking on each field and using the gray-shades bar-wheel at the top or on the "<" and ">" buttons placed at its sides; two vertical red lines will mark the corresponding positions on the graph. The ***Reset q range*** button will re-expand the q range.
The last row contains the scaling settings/commands:

- The *Scale to:* round checkboxes will scale the curves to either the *Minimum* (***default***) or the *Maximum* values among the selected curves.
- The *Save interpolated to target* checkbox ensures all q or t grids of the data will be identical by interpolating each curve to the minimum or maximum intensity curve. This is a convenience tool to allow curves generated on different grids to be put on a common grid.
- ***Reset scaling*** will return to the initial situation.
- ***Apply*** will perform the scaling.
- The ***Create scaled set*** will generated an actual set of scaled data which can be then saved as files.

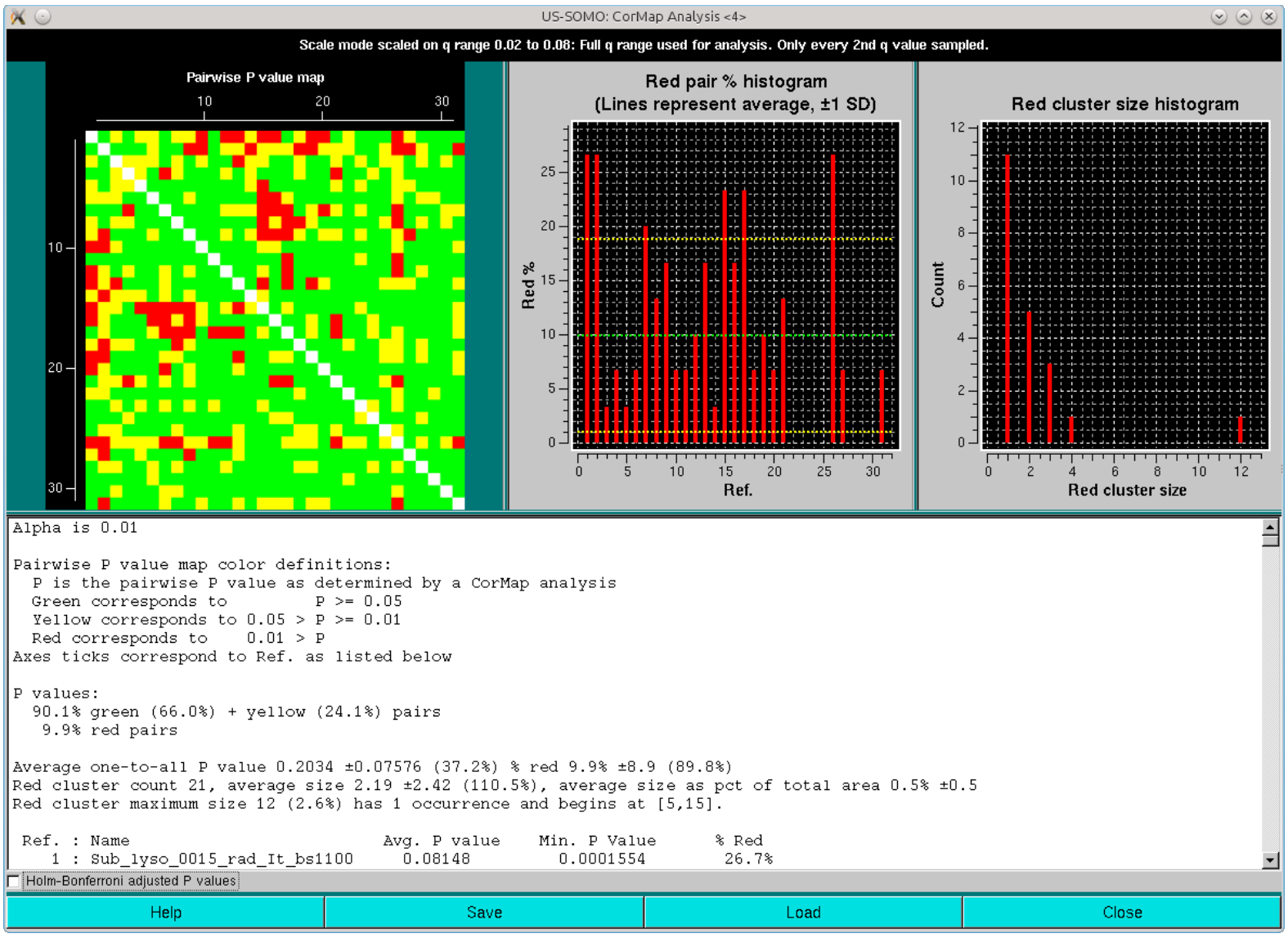
In the image below, the scaling has been performed on the indicated q range. The *Messages* panel reports the statistics of the scaling process as applied to each curve scaled on the one with the lowest intensity:



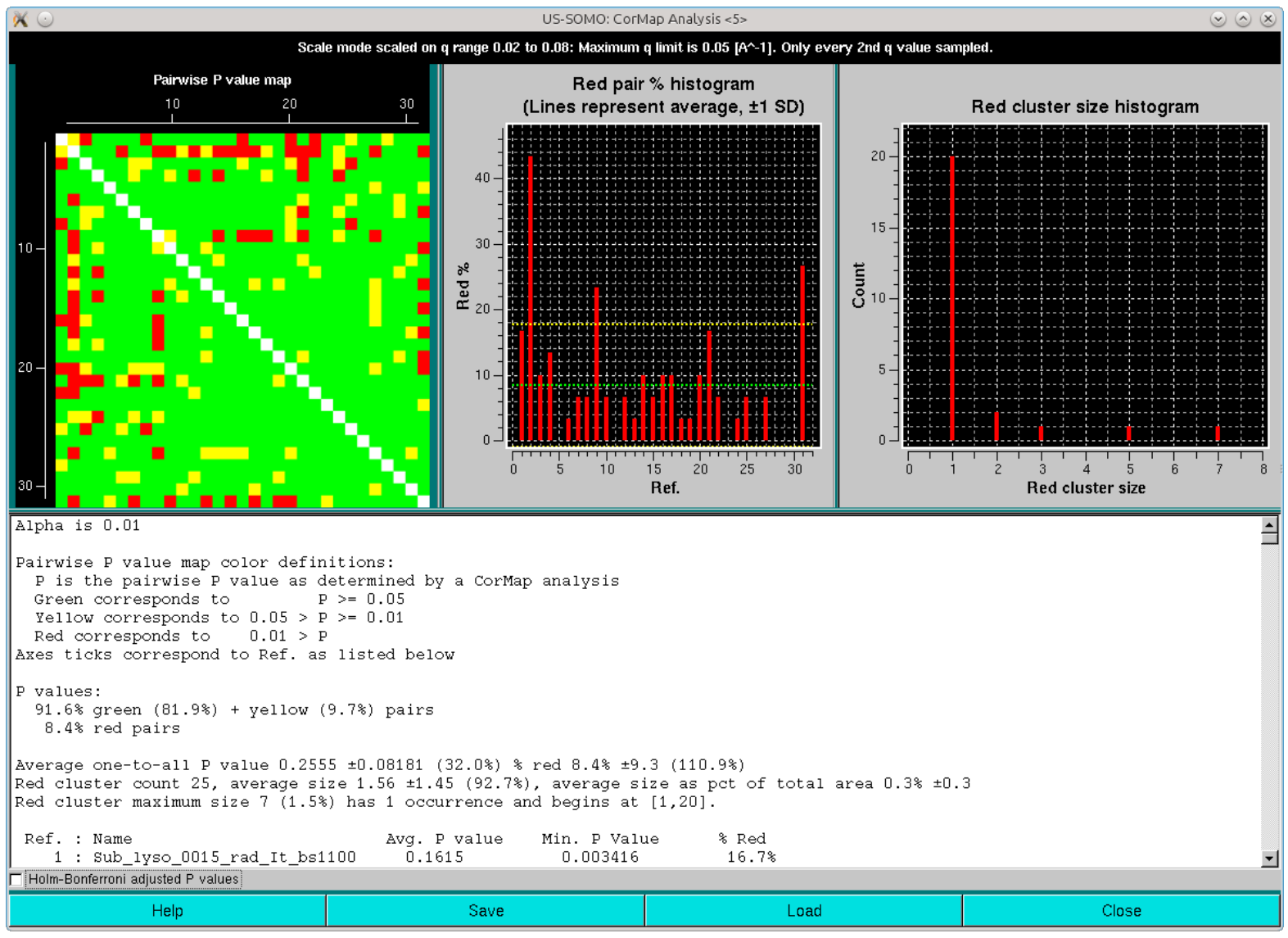
By pressing ***Residuals***, the residuals of the scaling operation are also shown. If the ***Scroll*** checkbox is selected, the scaled files can be examined one at a time, scrolling through them using the gray-shades wheel-bar or the "<" and ">" buttons placed at its sides. The file name of the data currently shown will appear below the gray-shades wheel-bar, as shown here:



A **CorMap** analysis can be also performed on the scaled set by pressing ***CorMap Analysis***. This will produce two pop-up outputs, one containg the CorMap analysis on the full *q*-range available:

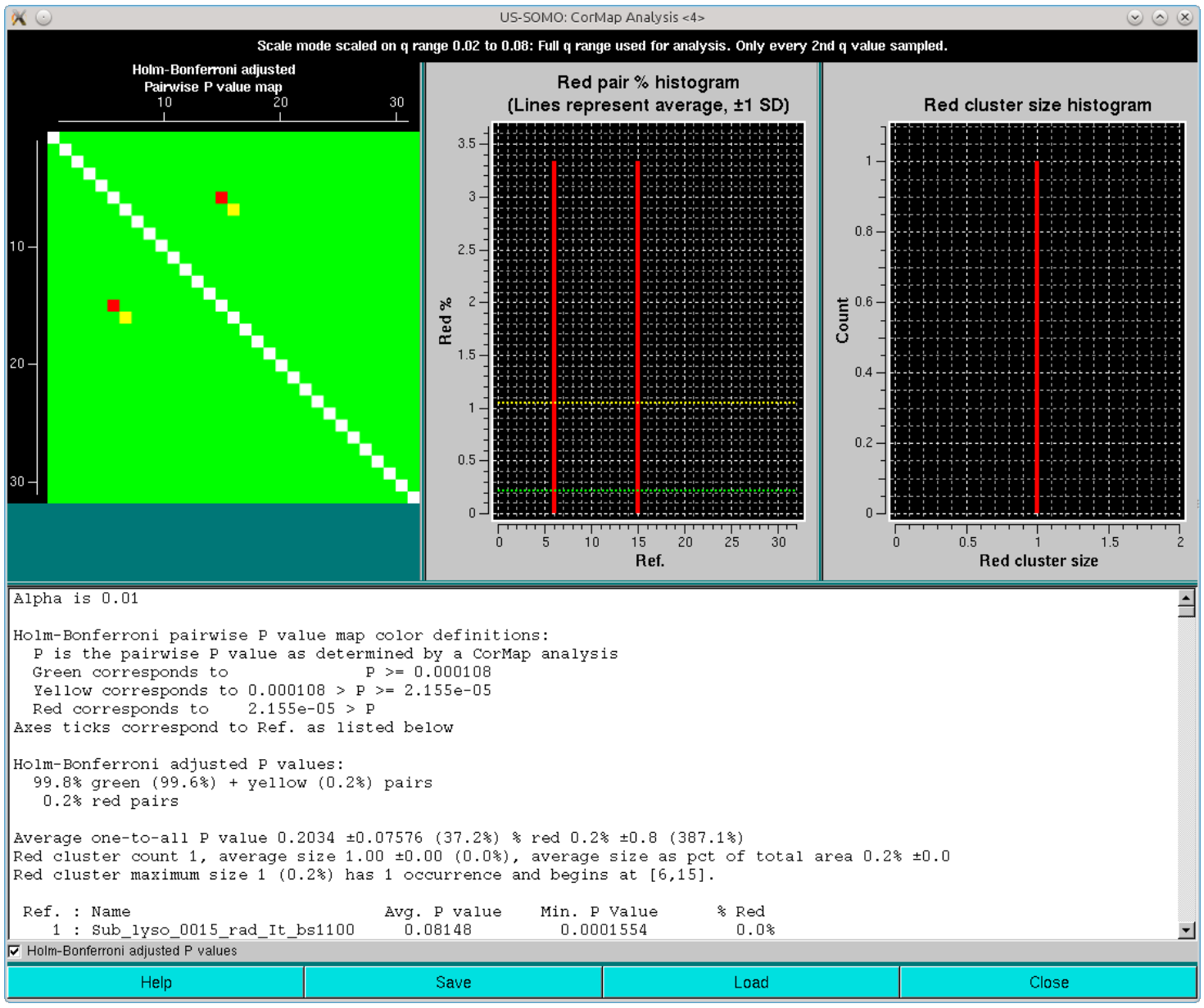


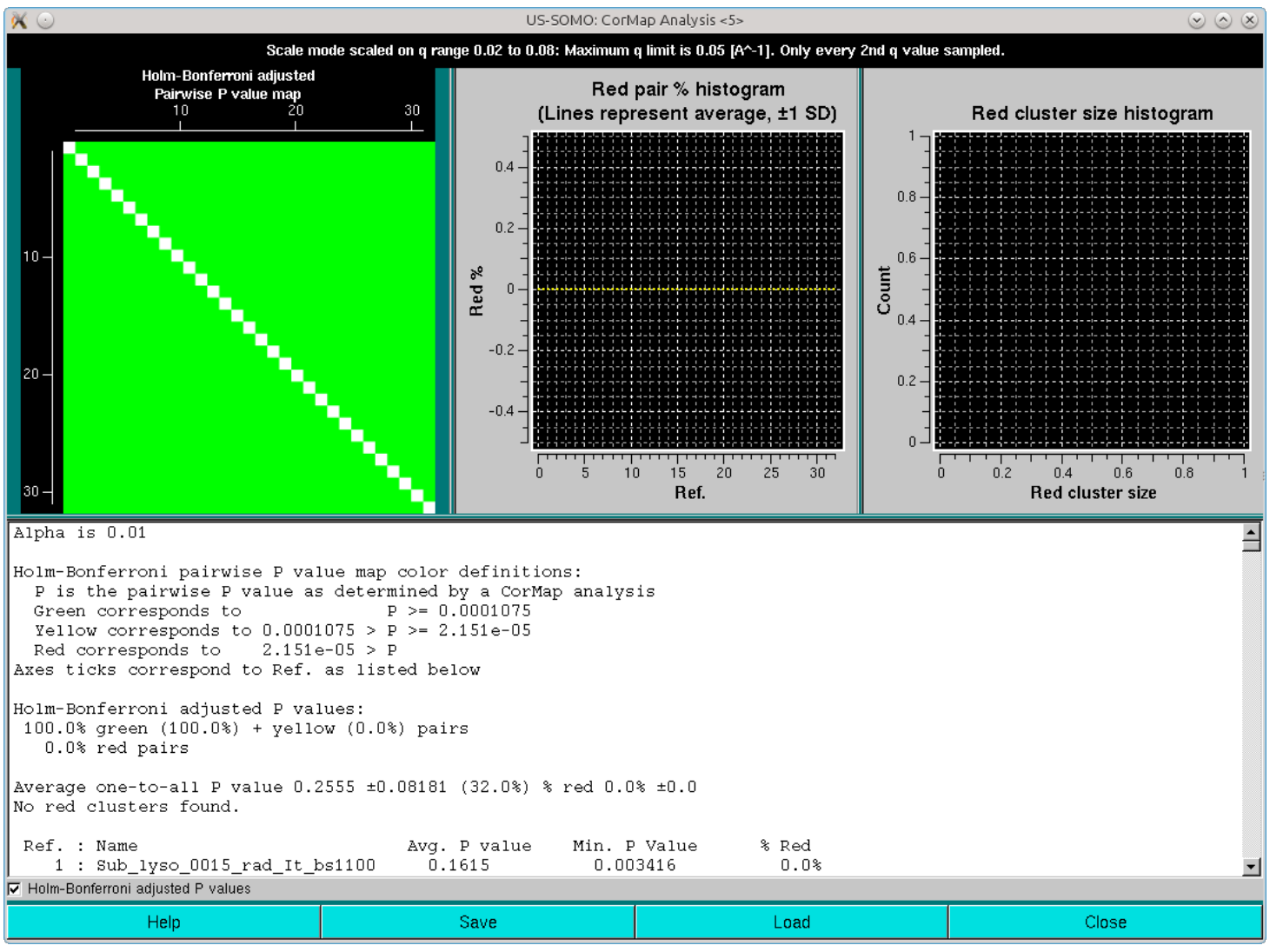
and the other limited to the q_{max} value presente in the ***Options*** (0.05 Å^{-1} in this case):



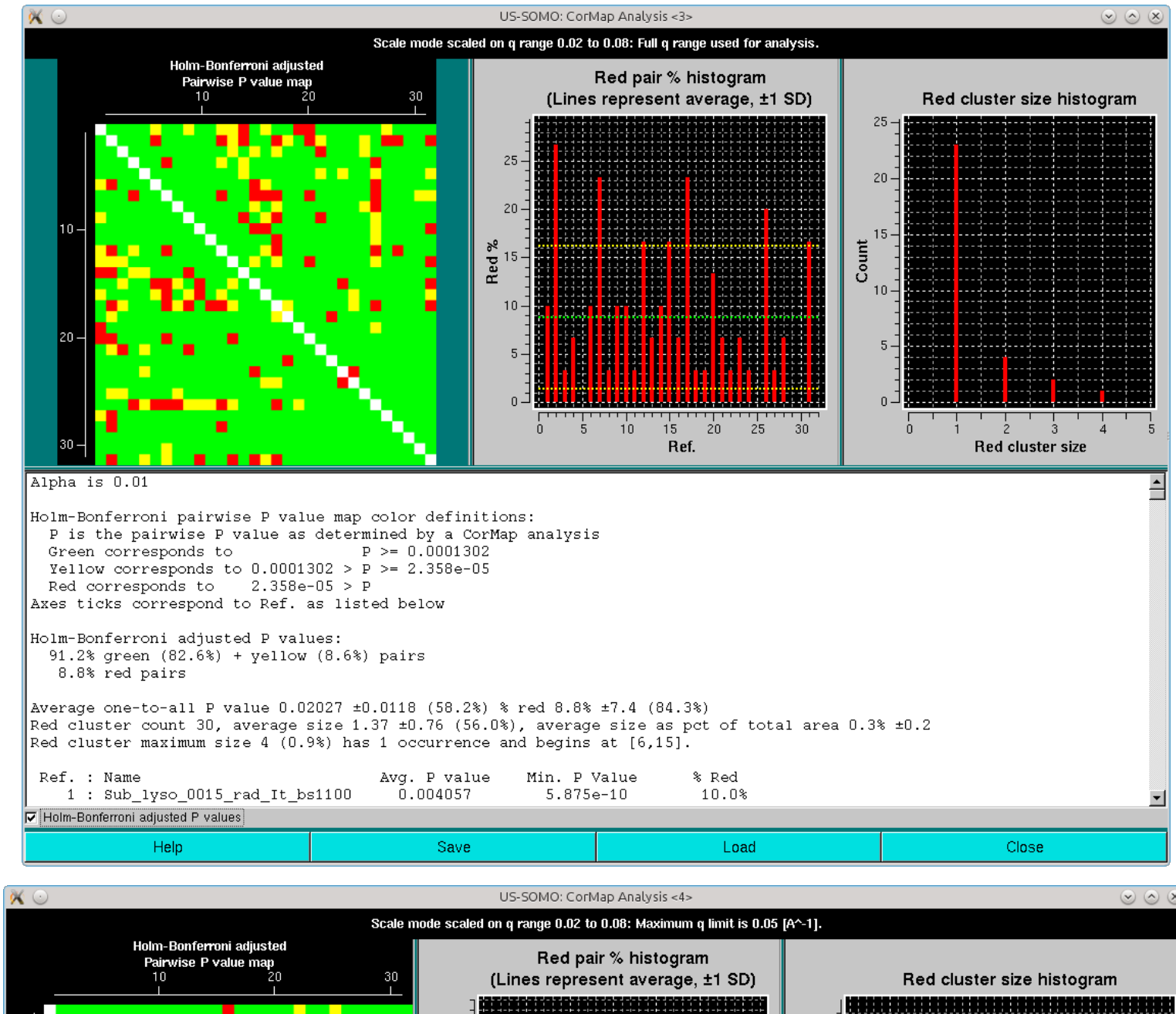
Note that one-every-other q-value has been utilized in this analysis. As can be seen, both the pairwise P-value map and the red pair % histogram indicate that with the exception of three frames of the ensemble (#2, #9 and #31, corresponding to frames #1101, #1108 and #1130), most are similar to all the others, with an overall ≈8-9% of red pairs. Limiting the lower q-values to 0.015 Å^{-1} reduced this overall % red pairs count to 6% (not shown), with frames #1101 and #1130 still being substantially different from all others.

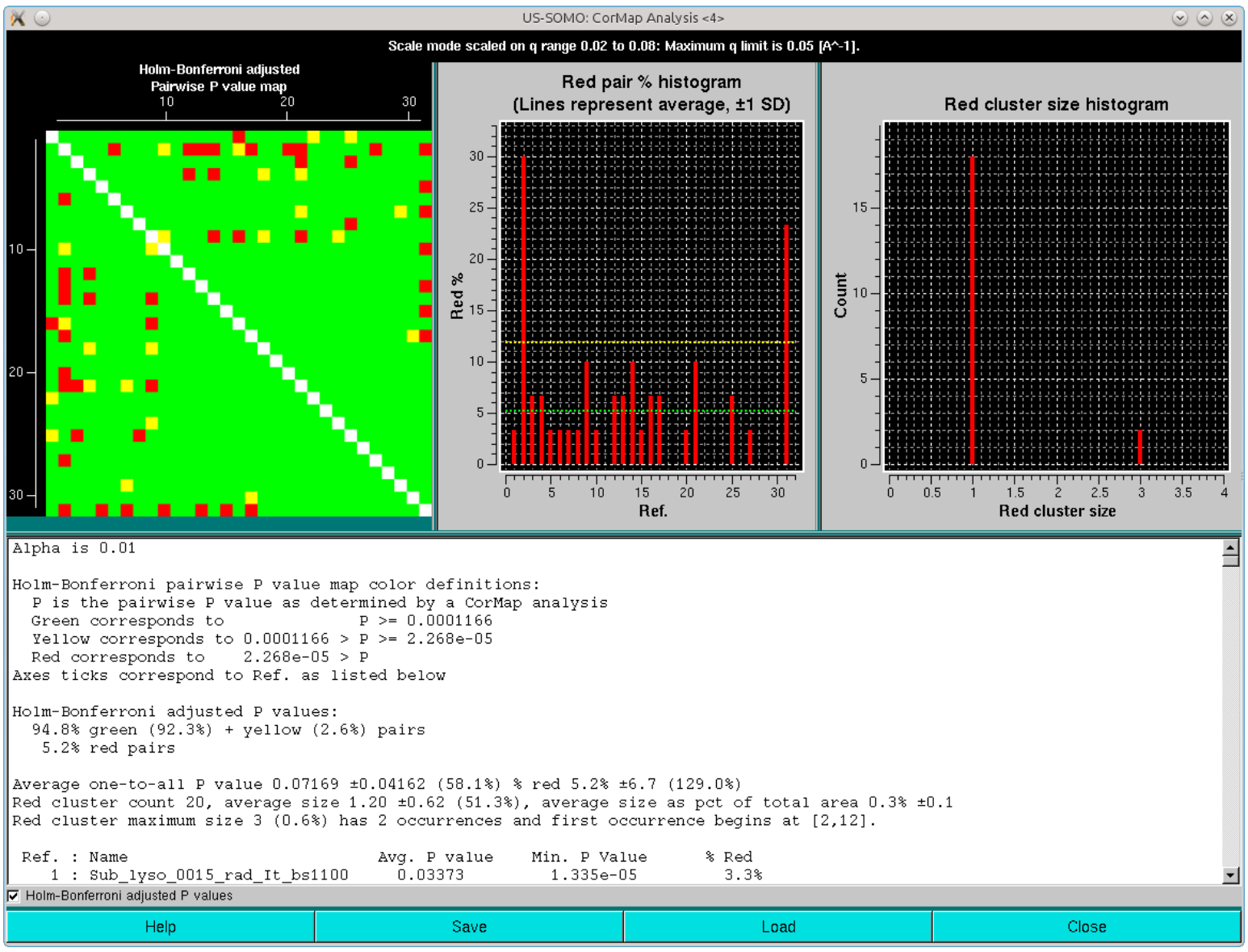
If the *Holm-Bonferroni adjusted P values* checkbox is selected, these will be the results:





Clearly, the combination of the one-every-other *q*-value sampling and the Holm-Bonferroni adjustment is over-permissive for this dataset. If the analysis is repeated without the sampling, these are the results:





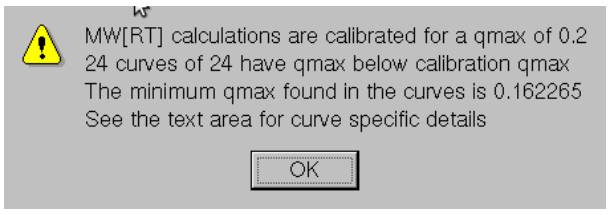
In this case, similar results are obtained as with the one-every-other q-value sampling and no Holm-Bonferroni adjustment.

Pressing ***Cancel*** will completely exit from the ***Test I(q)*** mode.

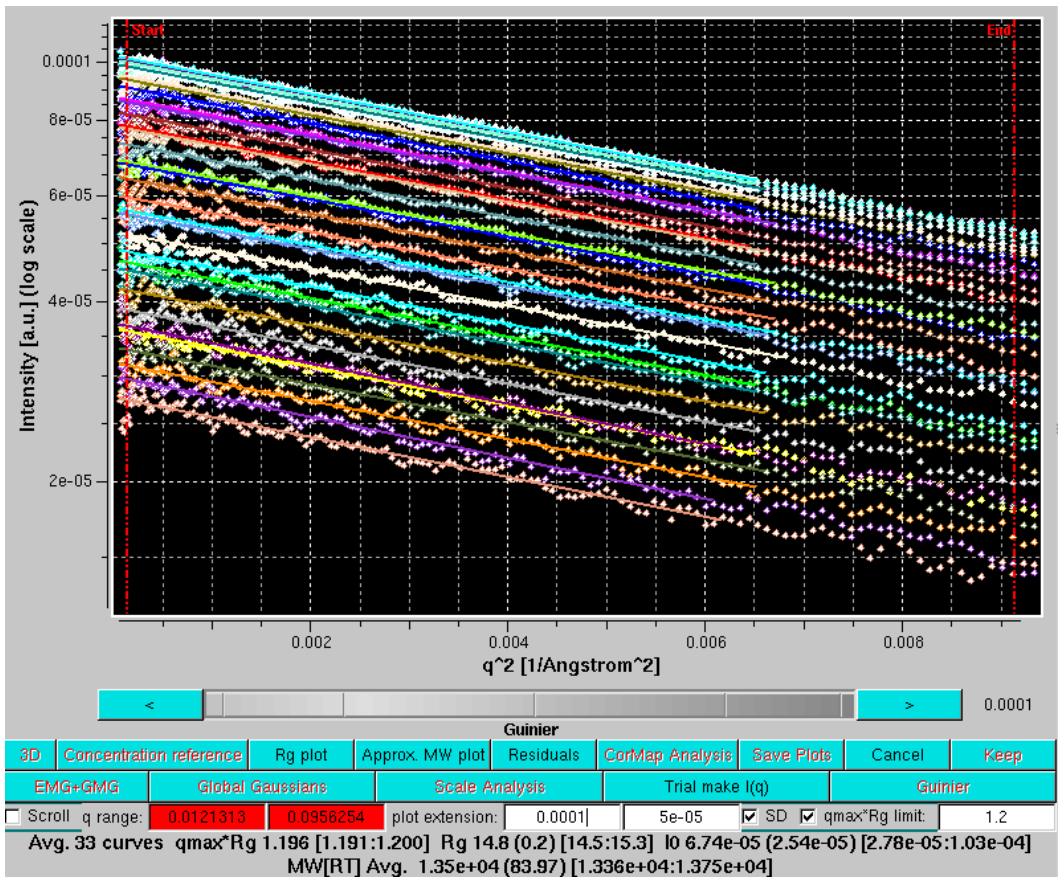
Pressing ***Trial make I(q)*** will instead exit from the scaling mode only.

Pressing then the ***Guinier*** button will call the other test function available in the ***Test I(q)*** mode:

If the maximum q value for the currently examined $I(q)$ vs. q dataset does not reach the ***MW[RT] q_{max} cut-off*** value present in the ***Options*** panel for the Rambo and Tainer approximate molecular weight calculation method (see here), a warning will appear:



Pressing ***Ok*** will allow to proceed, showing the *Guinier* mode of the ***Trial make I(q)*** panel (in the case examined, the pop-up alert dit **not** appear, as the q_{max} utilized was >0.2 Å^{-1}):



The lowermost row now carries the tools necessary to perform a Guinier analysis on the back-generated temporary $I(q)$ vs. q curves:

- The *q range:* fields (red background) indicate the q range currently selected. By clicking on each one and using the gray-shades bar-wheel at the top or the " <" and ">" buttons placed at its sides, the q-range can be changed.
- The *plot extension:* fields provide a q range extension of the Guinier and Residuals plots, allowing additional q range before (left box) and after (right) the Guinier range to be shown. By clicking on one and using the gray-shades bar-wheel at the top or the "<" and ">" buttons placed at its sides, the q-range extension can be changed.
- The *SD* checkbox controls if the linear regressions for the Guinier analyses are carried out with SD weighting (***default: checked***).
- The *qmax*Rg* checkbox and its associated field enable a sequential routine which examines, starting from the higher q value utilized in the regression (q_{max}), if the product $q_{max}*R_g$ is ≤ the limit set in the apposite field (the ***default*** value is set in the ***Options*** panel). If the products exceeds the limit, that q_{max} value is dropped, and the regression is repeated utilizing the next lower q value. The test is repeated until a q_{max} value satisfying the $q_{max}*R_g \leq$ set limit is found. It is highly advisable to **deselect** this checkbox before manually changing the *q range* fields, and to **re-select** it afterwards, otherwise the program will continuously repeat the $q_{max}*R_g$ routine while the limits are manually moved.

At the bottom of this window region, the averages values for all the curves are reported, and they include:

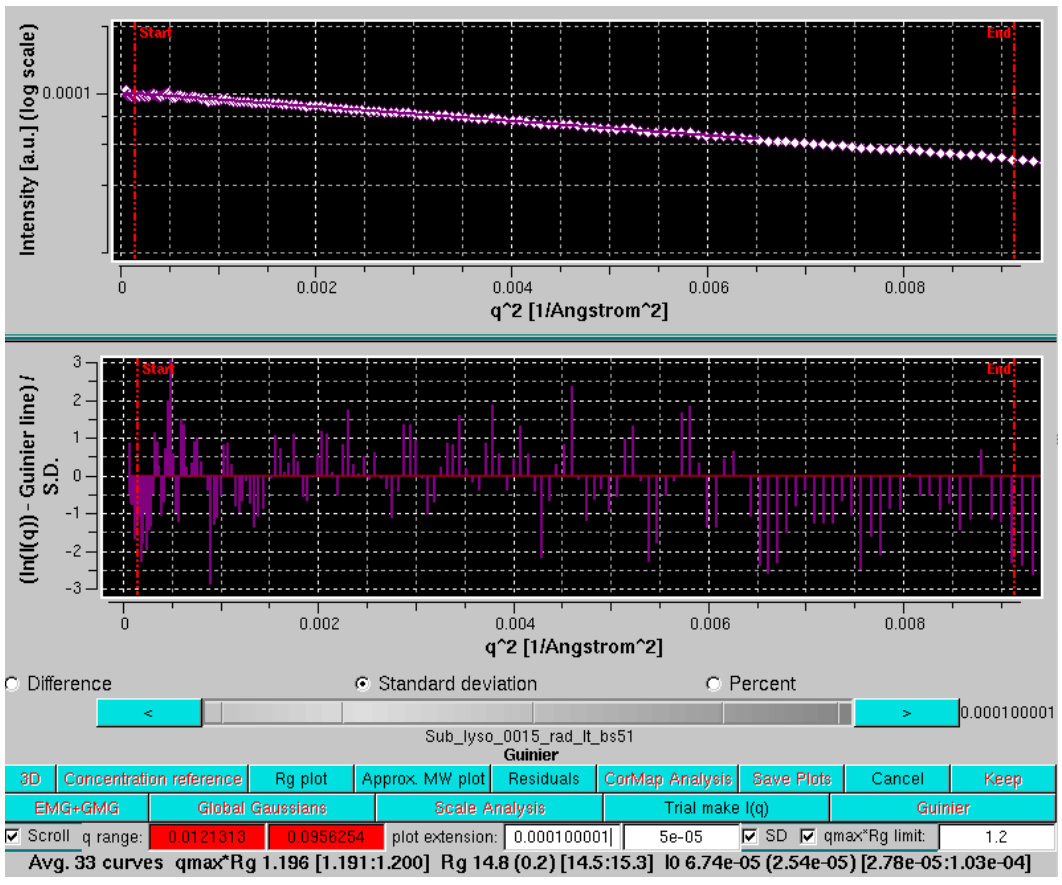
- The number of curves used (***Avg. xx curves***);
- The average ***qmax*Rg*** (where ***qmax*** is the maximum q value included in the Guinier analysis), with the range in square brackets;
- The average ***Rg***, with the SD in round brackets and the range in square brackets;
- The average intensity extrapolated at 0 scattering angle ***I0***, with the SD in round brackets and the range in square brackets.
- In the line below, an approximates molecular average molecular weight, ***MW[RT]*** is reported, together with its SD (in round brackets) and range [in square brackets], based on the method of Rambo and Tainer (Accurate assesment of mass, models and resolution by small-angle scattering. Nature 496:477-481, 2013). If the available q-range is below that required by the method, a warning message appears in the *Messages* area.

The regression data for each individual curve are reported in the *Messages* window.

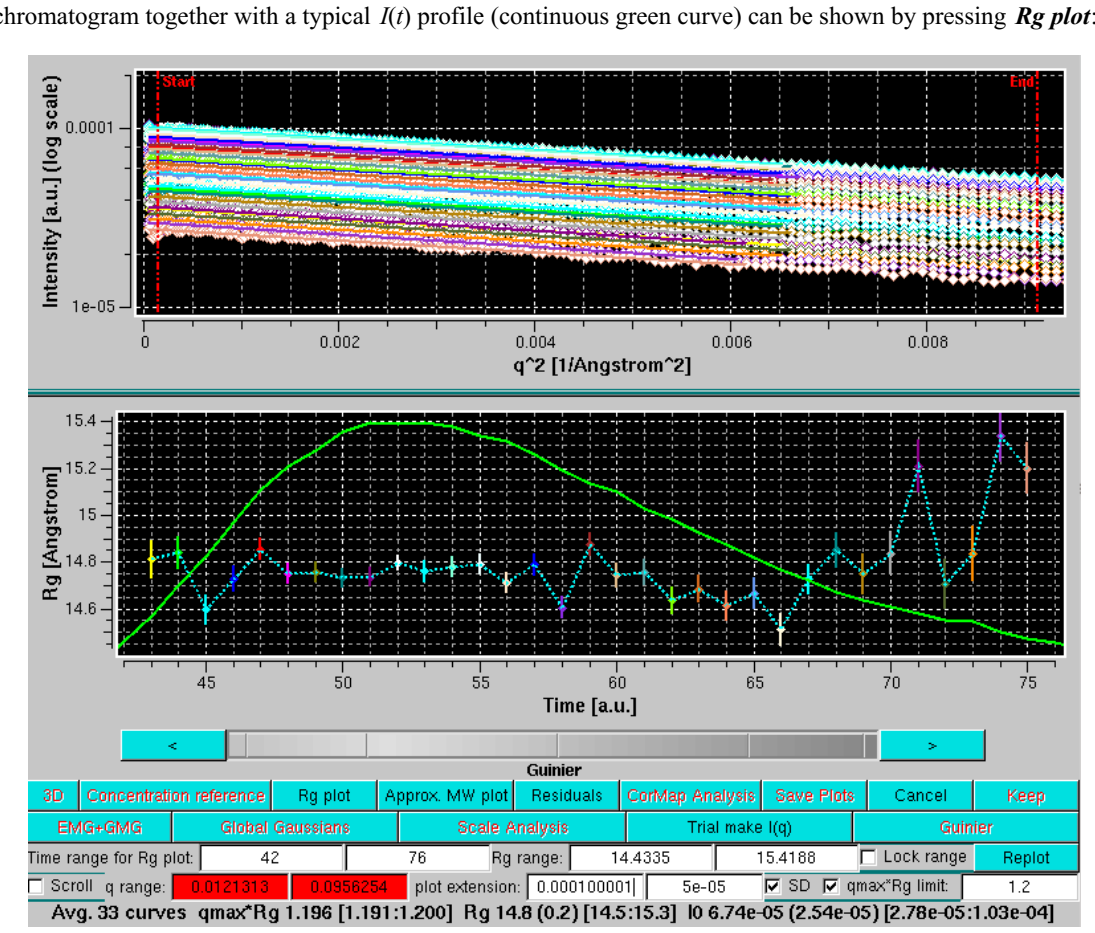
The residuals of each linear regression can be seen by pressing ***Residuals***:

Note how the average *Rg* recovered for this extended set, 14.8±0.2 Å, compares well with the 15.0 Å that can be calculated from the lysozyme average NMR structure (1E8L.pdb) using the WAXSiS server (http://waxsis.uni-goettingen.de/).

As with the scaling option, every individual Guinier plot can be visualized by selecting the ***Scroll*** checkbox and using the gray-shades wheel-bar:



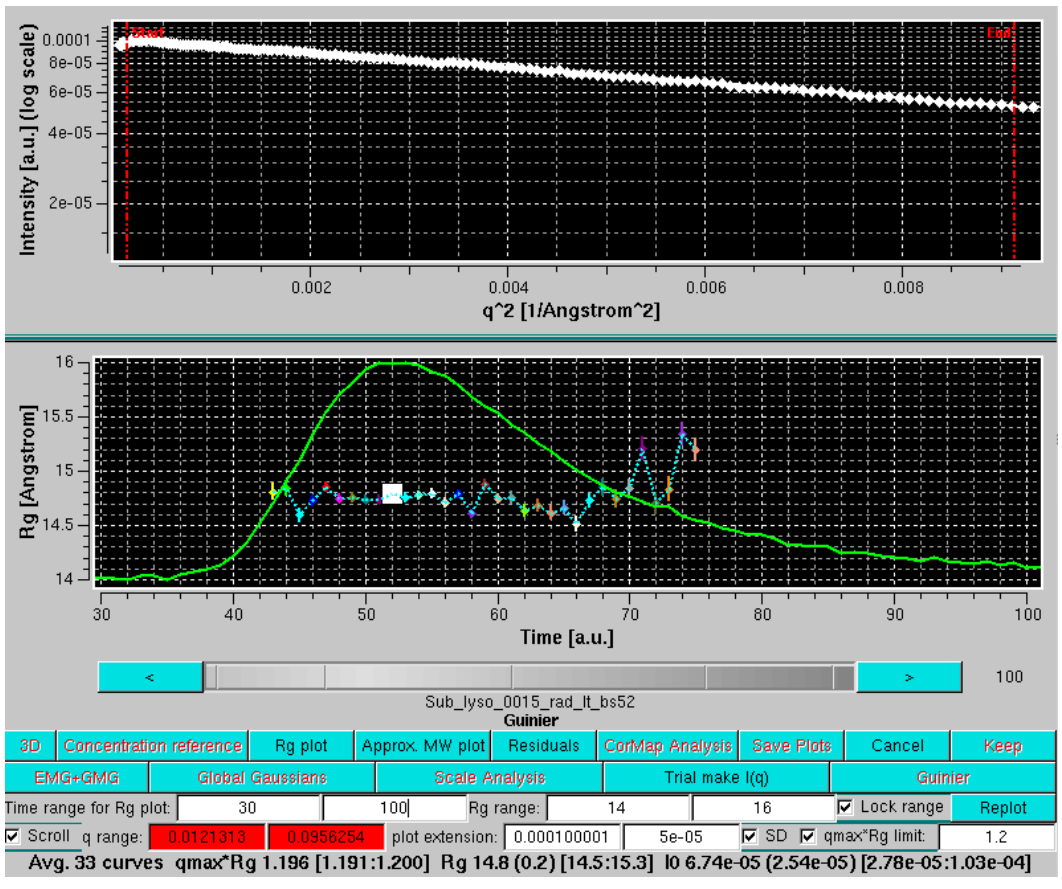
At this point, a plot of the *Rg* values across the chromatogram together with a typical *I*(*t*) profile (continuous green curve) can be shown by pressing ***Rg plot***:



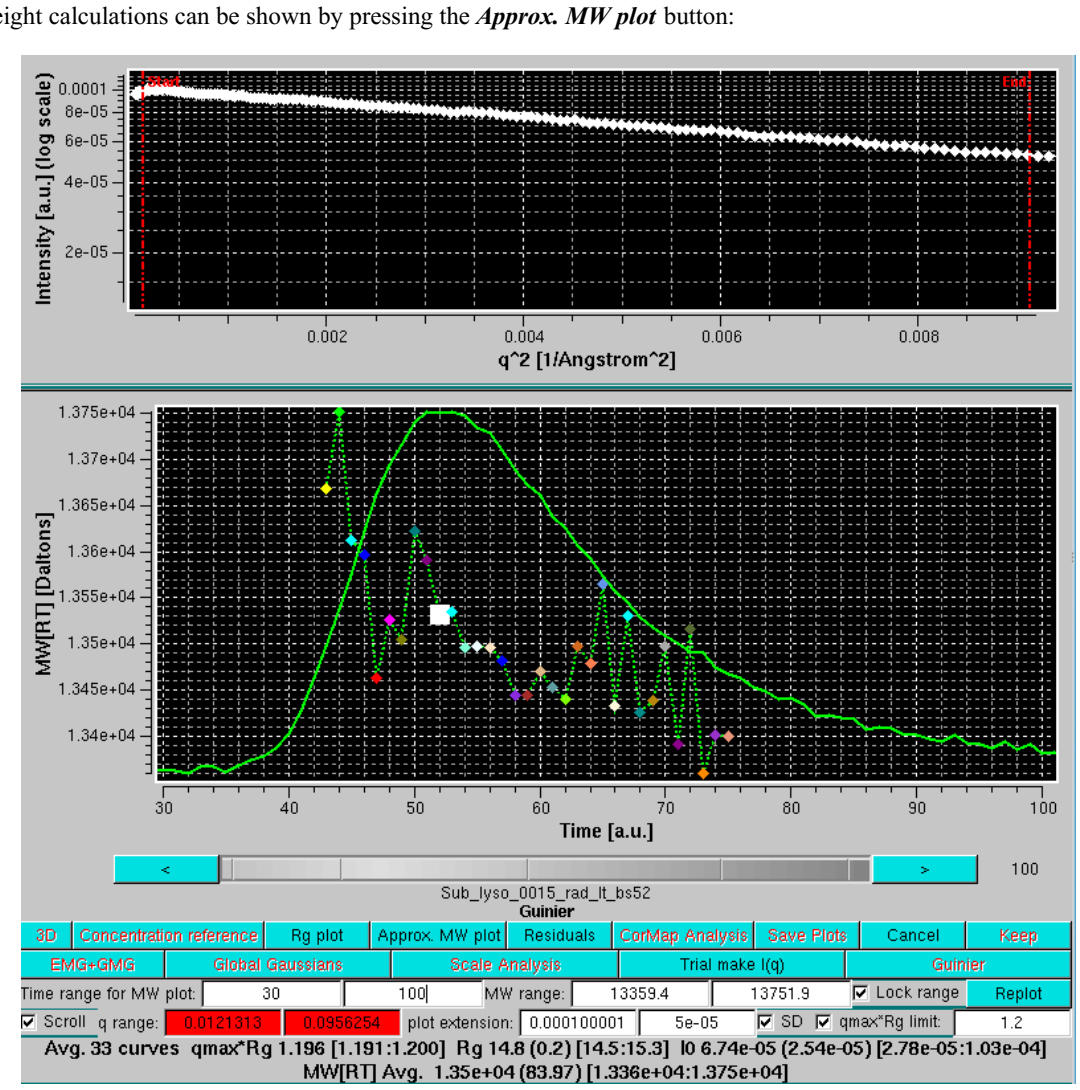
A new row will appear below the graphics window, with these fields:

- *Time range for Rg plot:* with its two fields. Clicking on each one and using the gray-scale bar-wheel at the top or the " <" and ">" buttons placed at its sides will allow to expand or restrict the Time axis (by default, the original time range is shown).
- *Rg range:* with its two fields. Clicking on each one and using the gray-scale bar-wheel at the top or the " <" and ">" buttons placed at its sides will allow to expand or restrict the R_g range. The default is governed by the actual R_g with error bars range with an additional empty space of 2.5% of the total range added to the lower and upper R_g limit.
- The *Lock range* checkbox if selected will maintain the actual shown R_g range. Pressing the ***Replot*** button will replot the data keeping the selected R_g range and maximizing the *I*(*t*) profile within the selected time range limits.

The scroll capability can also be activated in this mode, and the currently selected Guinier plot will be highlighted in the R_g plot:



Likewise, plots of the approximate molecular weight calculations can be shown by pressing the ***Approx. MW plot*** button:



Pressing ***Test I(q)*** button will bring back the main options of this utility.

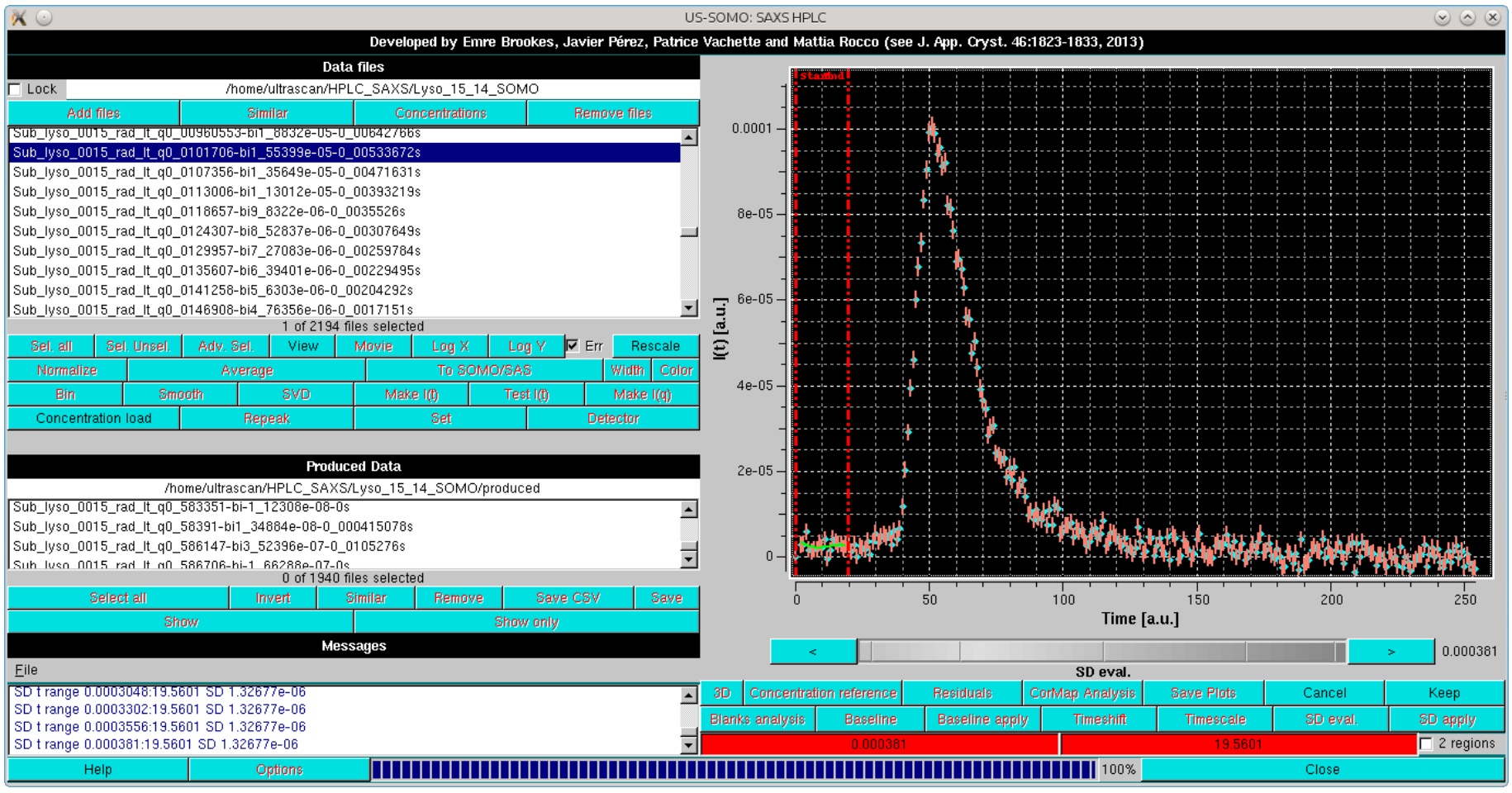
Pressing the ***Cancel*** button will exit the ***Test I(q)*** utility.

If Gaussian analysis is not required, a series of *I(q)* vs. *q* frames can be re-created at this stage from the baseline-corrected data by pressing the ***Make I(q)*** button.

Another new feature present in this release of the ***US-SOMO HPLC-SAXS*** module is an alternative way of estimating the errors associated with the SAXS data. This might become useful if no errors have been already associated with the experimental data, or if their reliability is questionable.
The method is based on the assumption that the fluctuations of the signal at the baseline level are a good representation of the error associated with the data an any other point along each *I(t)* vs. *t* chromatogram. Therefore, by estimating the average fluctuations in flat regions of the chromatogram, a **constant** SD value can be associated with **every** datapoint in that particular chromatogram. Obviously, different chromatograms will have different values for their respective constant SD.

Upon selecting a single *I(t)* vs *t* chromatogram, the actual SD associated with it can be visualized by selecting the ***Err*** checkbox in the ***Data files*** commands section, as shown in the image below:

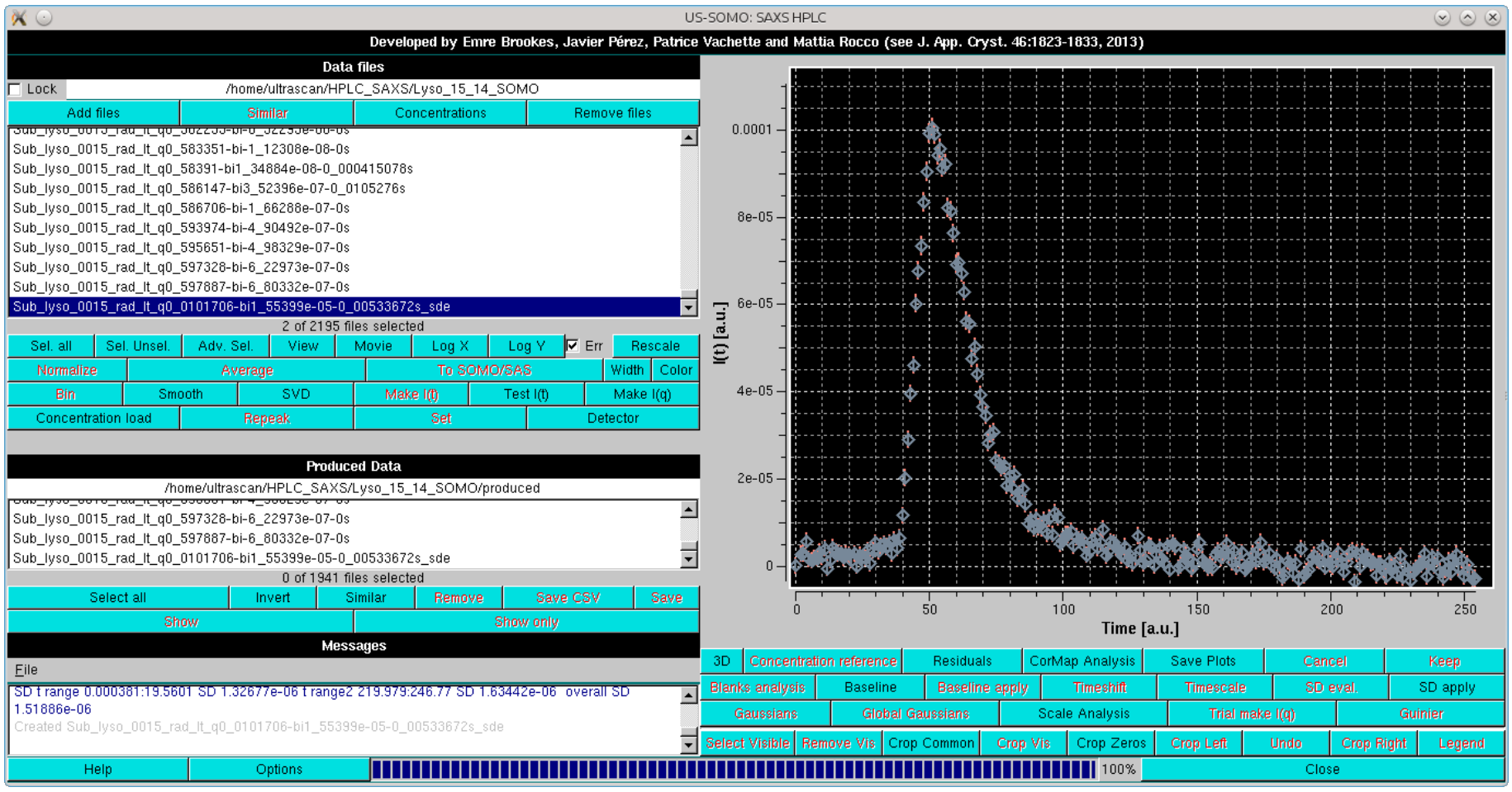
By pressing the ***SD eval.*** button, two vertical red lines will be superimposed on the initial region of the chromatogram, and two red-background fields controlling their position will appear in the bottom row of the commands below the graphics window:



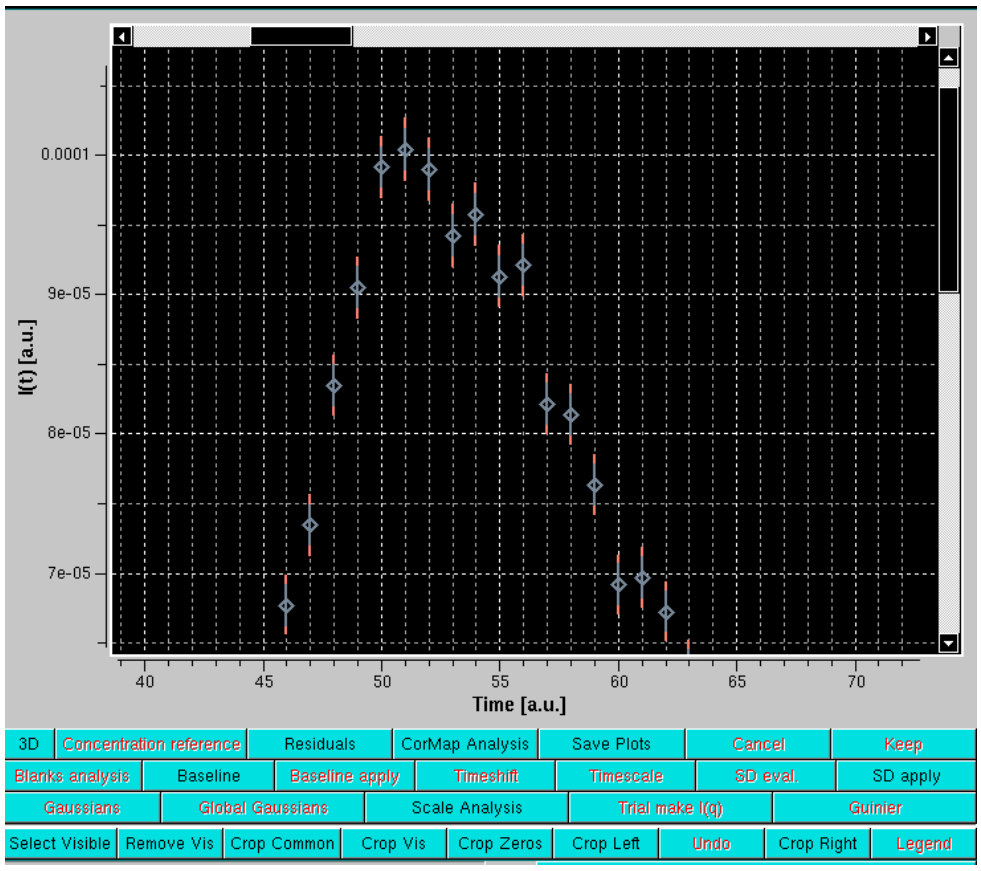
An additional checkbox, ***2 regions***, if selected will duplicate the vertical lines and their associate fields (colored magenta this time), allowing to utilize more than one flat region for the SD estimation:

The SD evaluation is carried out by fitting the data included between each zone with a 3rd degree polynomial, and taking the RMSD of the fit as the SD. If two regions are chosen, the final SD will be an average between the values computed from each region.

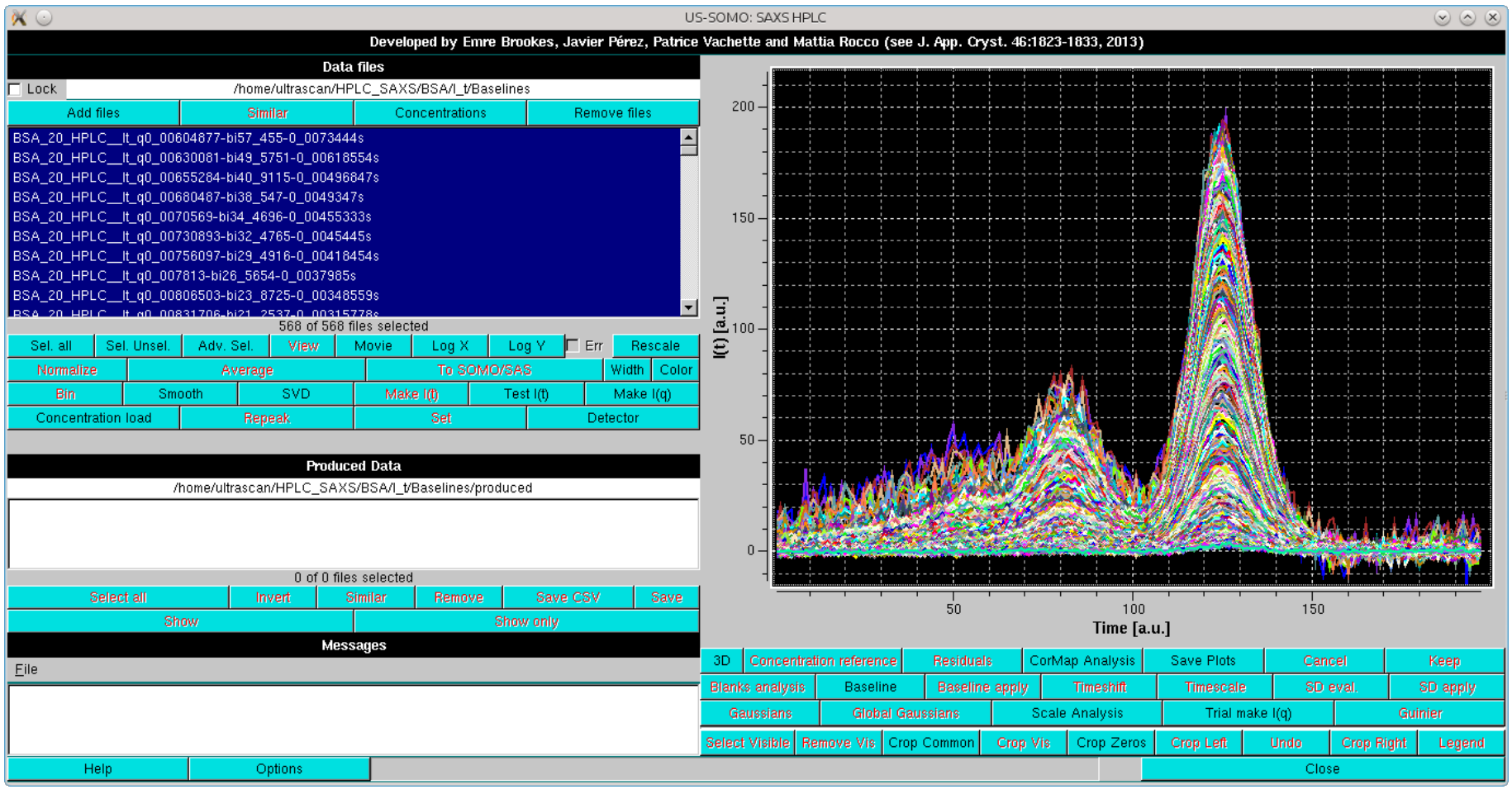
After adjusting the zone(s) for the SD evaluation, pressing ***Keep*** will accept the values, and the ***SD apply*** button becomes available. Pressing it will apply the SD calculation to any selected chromatogram. In the example below, the same chromatogram is plotted twice, with the old (salmon) and new (slate gray) SDs:



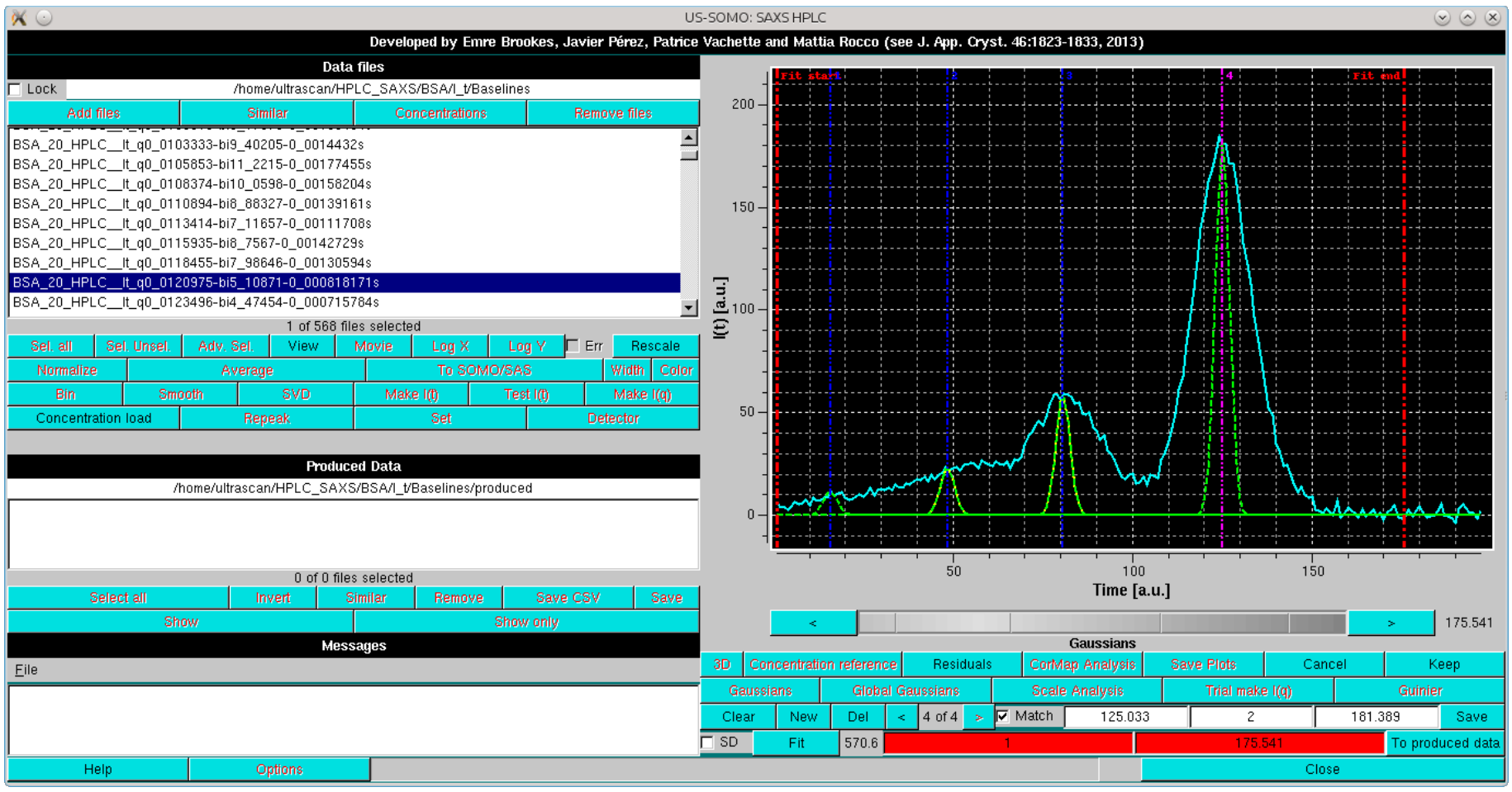
A blow-up of the main peak region highlights how the two SDs are very close to each other, demonstrating that the assumptions taken for this alternative SD evaluation produce SDs which are very similar to the ones that have been associated to the original SAXS data using a Poisson distribution. The main difference is that the original SDs vary slightly with the intensity for each *I(t)* vs. *t* chromatogram, while the baseline fluctuations method produces constant SD values for each *I(t)* vs. *t* chromatogram.



Gaussian decomposition of not baseline-resolved peaks is another utility present in the **US-SOMO *HPLC-SAXS*** module. Decomposition with symmetrical Gaussian functions will be first described using a bovine serum albumin (BSA) SEC-SAXS run using two 7.8 × 300 mm ID columns packed with hydroxylated polymethacrylate particles (TSK G4000PWXL, 10 µm size, 500 Å pore size, and G3000PWXL, 6 µm size, 200 Å pore size, Tosoh Bioscience, Tokyo, Japan) connected in series, protected by a 6 × 40 mm guard column filled with G3000PW resin (Tosoh). The data presented capillary fouling evidence, and were thus subjected to Integral Baseline correction (not shown).



Before proceeding to Gaussian analysis (whose theory can be seen here), a SVD analysis could be useful. In SVD analysis, the number of significant singular values in the decomposition should be equal to the number of components in the data, and thus to the minimum number of Gaussians required to accurately reconstruct the data (see here).



The baseline-subtracted data can be subjected to Gaussian analysis by first selecting a single chromatogram, and then pressing the ***Gaussians*** button. By default, the **US-SOMO *HPLC-SAXS*** module will consider symmetrical Gaussians, but distorted Gaussian functions are also availble and can be selected from the ***Options*** menu (see here). The choice must be done before starting the following procedure. An example of a data processing with non-symmetrical Gaussians is presented **here**.

On pressing ***Gaussians***, two new rows will appear at the bottom of the graphics window. If a previously-generated set of Gaussians was present or loaded from file, the Gaussians will show up under the peak(s) together with vertical lines indicating their centers.

Clear will remove currently-generated Gausssian, and allow to start a new analysis. If Gaussians had been loaded from file, the ***Clear cached Gaussian values*** button in the ***Options*** menu should be used.

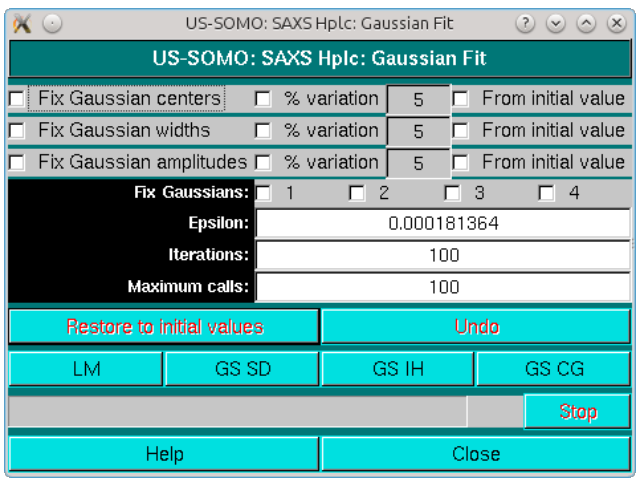
Each time the ***New*** button is pressed, a new Gaussian will be added (green colour), with pre-set center, width and amplitude shown in the three rightmost fields (additional fields will be present if distorted Gaussians functions are used; see here). By clicking on each field, and then using the gray-shades bar-wheel, each Gaussian can be adjusted to initialize the process (usually, only the centers need to be positioned under the peaks). If the *Match* checkbox is selected, the height of each Gaussian will be automatically adjusted so to match the height of the experimental $I(t)$ vs. t curve at the Gaussian current position.

Del will remove only the current Gaussian.

Clicking on the "<" and ">" buttons will toggle among the Gaussian present, whose identifying number is shown in the field between them. The active Gaussian is identified by a magenta vertical line positioned at its center, while blue lines are used for the others. The limits for the analysis of the chromatogram are shown with two vertical red bars, whose position is shown in the two red-background fields in the bottom row, before the ***To produced data*** button.

The *SD* checkbox controls whether the data associated std. dev. values will be used in the fitting procedure (***default: not selected***). It is recommended to select this checkbox only after a first round of fitting with the various algorithms provided has been performed, as the SD can only be used with the LM algorithm at the time of writing this Help section (April 2016).

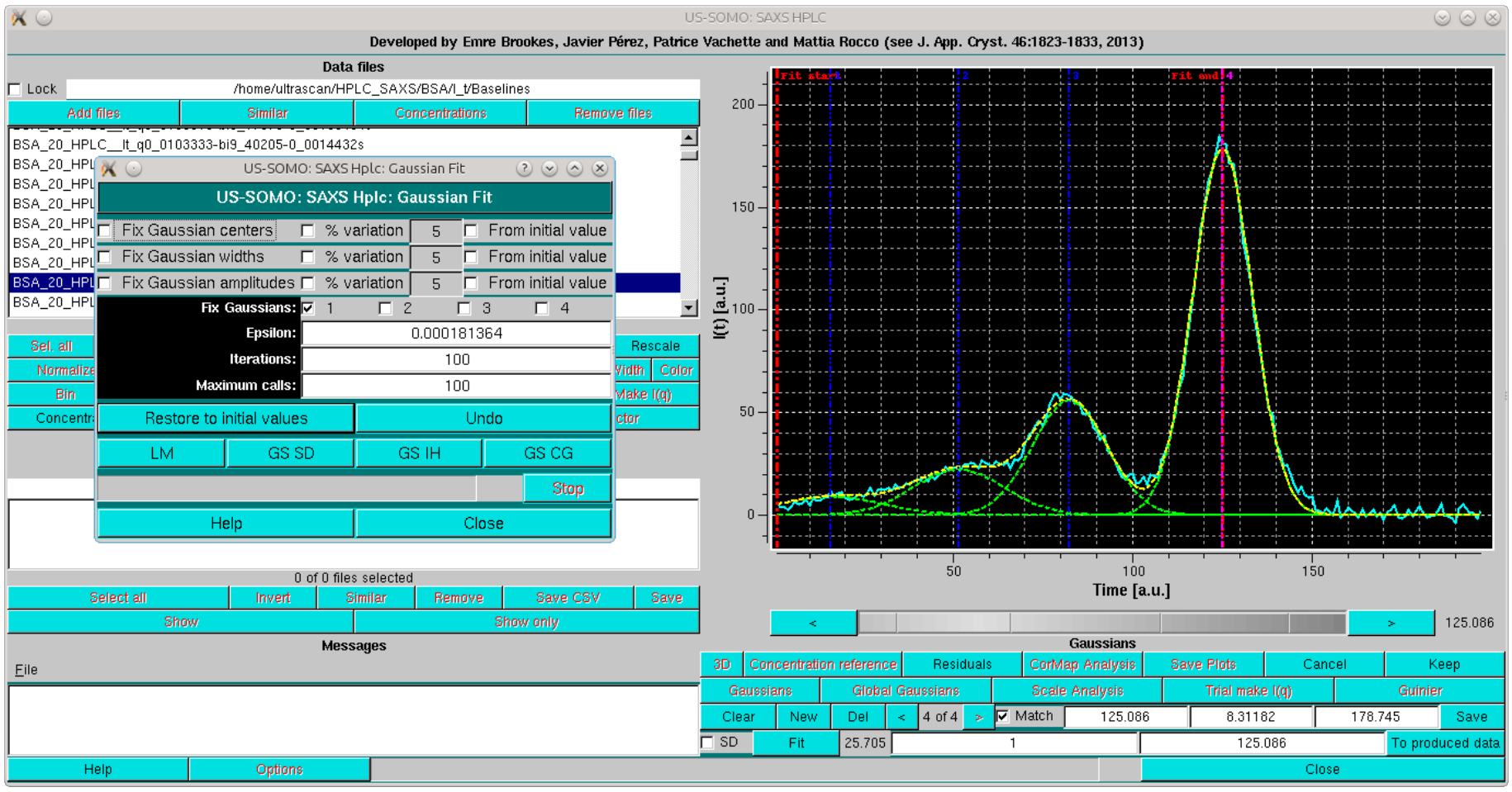
Once the initialization is completed, pressing the ***Fit*** button will bring up a window controlling the fit procedure, shown below:



See here for a description of the ***Fit*** module.

In the first cycle of iterations, it is best to keep the original centers fixed:

In the example shown, a not well-defined aggregates peak is present at the beginning, and an extended initial baseline is not present. If the first Gaussian is left free to adjust, it will expand too much to compensate for the missing initial baseline. Therefore, in such situations it is best to keep fixed the position of the first Gaussian:



A final round of fitting can then be performed using the SD and allowing a 5% variation on the Gaussian centers at each iteration, until a satisfactory fit of the main peak(s) is obtained:

If some datasets have missing or NaN values for one or more SD values, a pop-up menu will appear listing all the files presenting this problem, and with how many occurrences. The user can then select between three options: drop the datasets containing these non-defined SDs; drop just the frame (or time) point missing the SD(s); or not use SD weighting.

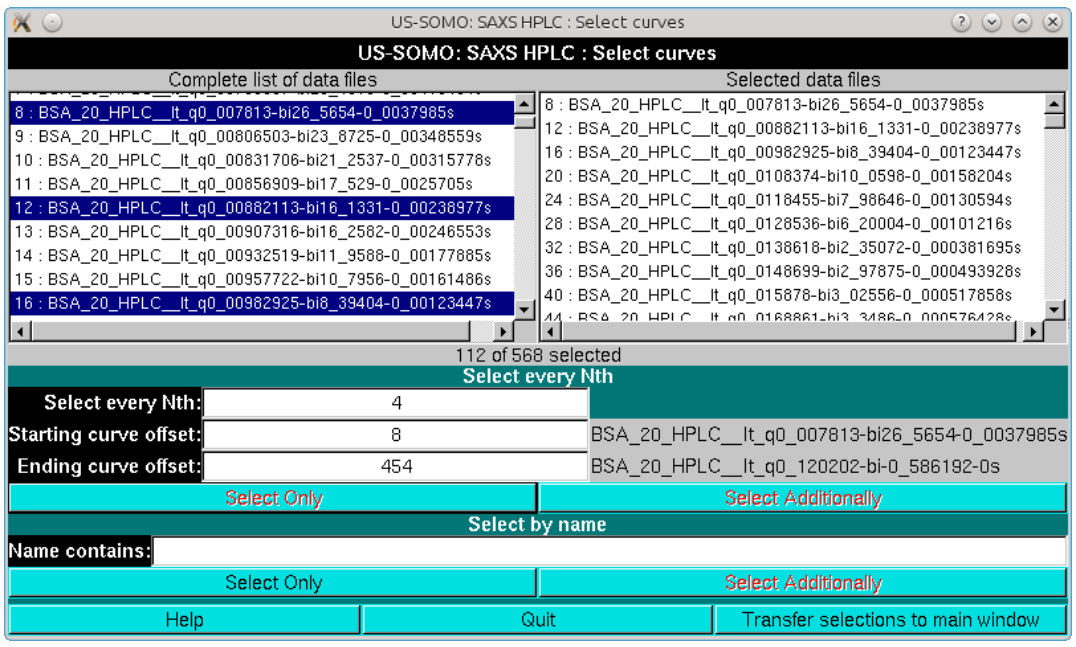
The global improvement of the fit can be also judged by the rmsd (*SD* checkbox not selected) or χ^2 (*SD* checkbox selected) value which is updated next to the ***Fit*** button. The residuals of the fit can be visualized by pressing the ***Residuals*** button, which will split the graphics window in two, and show a plot of the fit residuals below the main plot. The residuals plot can be removed by pressing ***Residuals*** again (see more below). In the example shown above, the residuals are weighted by the std. dev. associated with the experimental points (*SD* checkbox selected; a *By percent* residuals option is also available).

Once a satisfactory fit is reached, pressing ***Keep*** will accept the current Gaussians for further work. But to save the parameters of the current Gaussians in a file, the ***Save*** button has to be pressed before ***Keep***.

Cancel will cancel the operations and remove all the Gaussians.

Once an initial set of fitted Gaussians has been produced, it should be globally fitted to all chromatograms. However, performing this operation directly on all chromatograms can be very computationally intensive. For this reason, it is best to perform it on a subset of all chromatograms, and the global fit results are then propagated to all remaining chromatograms. **Importantly**, in the global fitting procedure the centers and widths of each particular Gaussian are optimized so to be the same across all chromatograms, and only the amplitudes are then fitted.

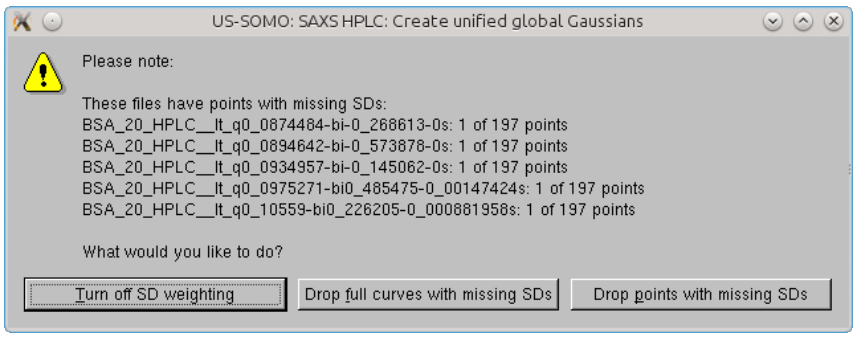
To select a subset of data, the ***Select*** button is pressed, which will open the pop-up selection panel (see image below and here).



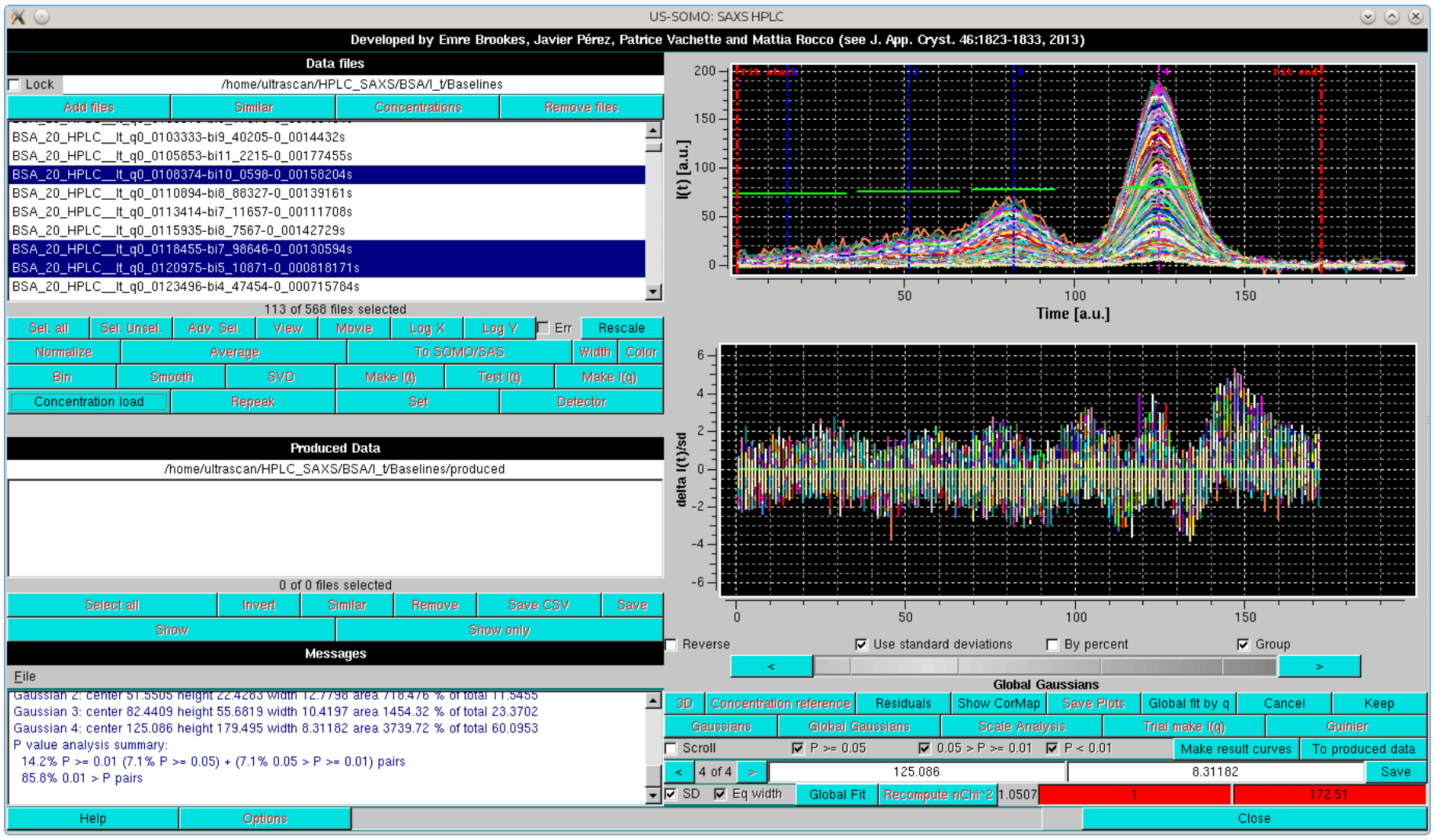
It is advisable to perform the global fitting avoiding the very first few low-q, very noisy, and the last high-q, very low signal $I(t)$ vs. t chromatograms. In the example we are illustrating, we start from chromatogram # 8 (q = 0.007813 Å^{-1}) and select every 4 chromatograms up to # 454 (q = 0.12020 Å^{-1}). The $I(t)$ vs. t chromatogram on which the initial set of Gaussians was optimized is also included (***Select Additionally*** button). Pressing ***Transfer selection to main window*** will close the pop-up window and the selected files will be shown in the main ***HPLC-SAXS*** module graphics window:

The ***Global Gaussians*** button is now available. Pressing it will simply find the amplitudes best fitting all the selected chromatograms based on the centres and widths found on the initial chromatogram. This operation has to be performed before the global fit.

If datasets having points with missing or NaN std. dev. values are found, a pop-up panel will appear:



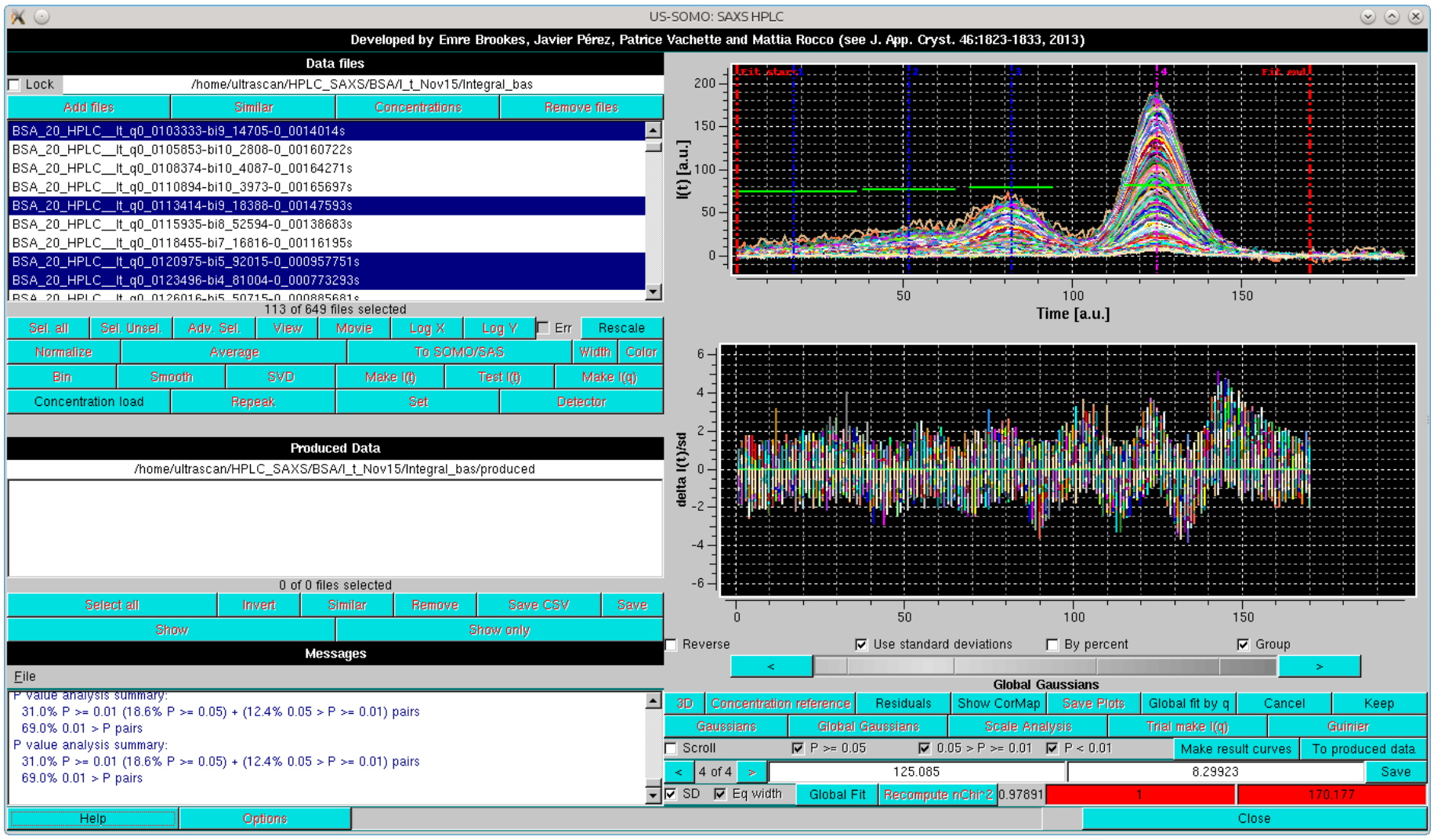
When just a few problematic points are found for each file, the *Drop points with 0 SDs* option can be safely used. The global Gaussians operation will then be completed:



In the image above, the ***Global Gaussians*** results on the ***Nth*** selected files are shown, together with the grouped fit residuals. The common centers and widhts, not optimized but just based on the initial chromatogram fit, are displayed in the graph as vertical and horizontal bars, respectively. Note that the *Residuals* plot *x*-axis scale was manually optimized (right-click on the scale) to make it comply with the fit limits in the top panel.

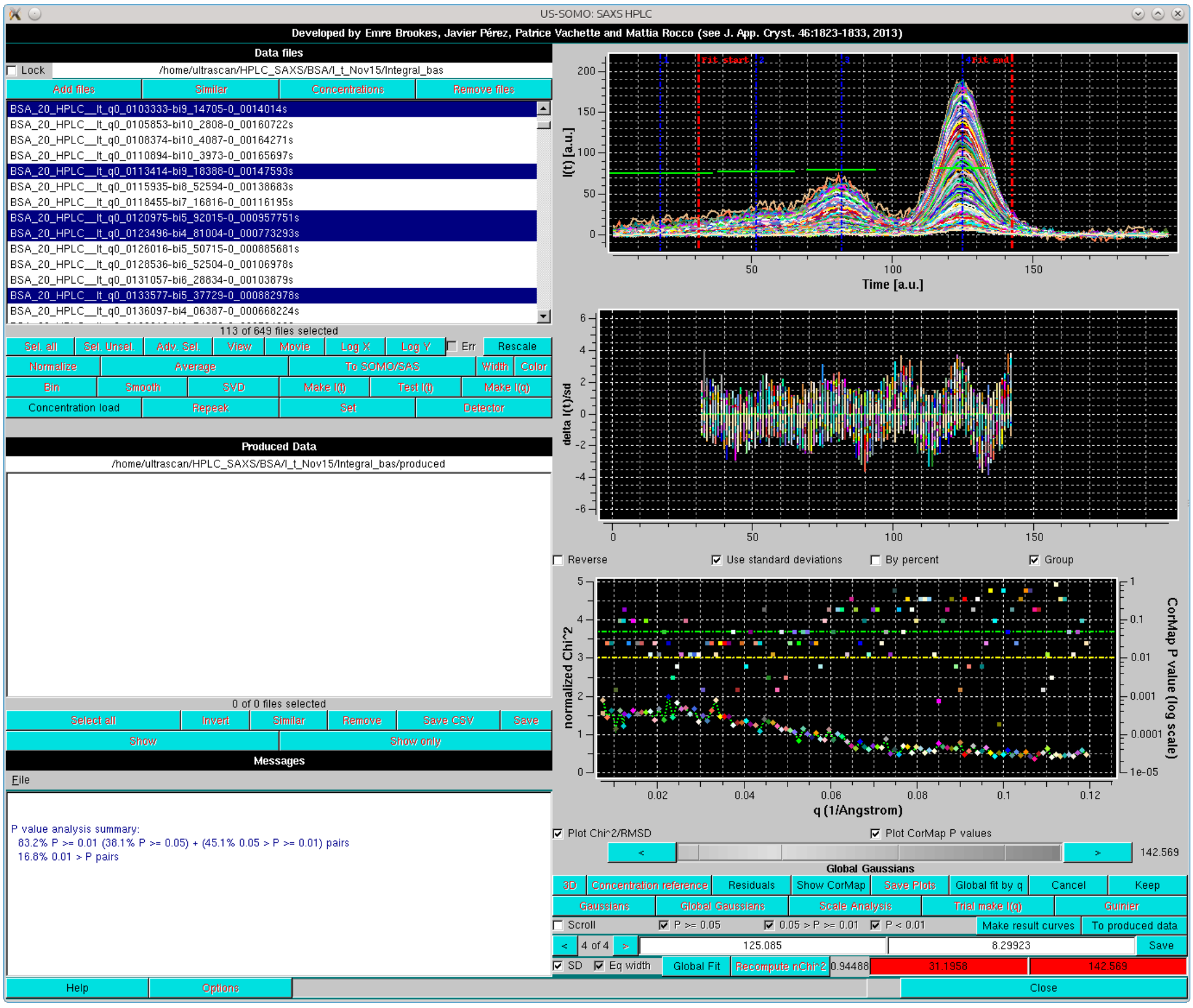
Save will save the resulting Gaussians to the current selected directory, with extension -gauss.dat for symmetrical Gaussians of single files and -mgauss.dat for Gaussians of multiple files. For distorted Gaussians, the extensions will be -mgmg.dat, -memg.dat, and -memggmg.dat for the GMG, EMG and EMG+GMG Gaussians, respectively.

Global fit, which becomes available instead of the ***Fit*** button once a series of chromatogram is selected and after at least an initial set of Gaussians is generated/loaded, can now be used to optimize all the centres and widths of each Gaussian along all the chromatograms to common values for each family of Gaussians. The operation is controlled by the same pop-up **Fit** panel as for the single chromatogram case (see here), but only the LM method is currently (April 2016) available. In this example, it is best to first perform a global fit round keeping the Gaussians 1, 2 and 4 fixed, and then perform a second round leaving all parameters free.



In the image above, the results of the ***Global fit*** are shown together with the grouped fit residuals. Furthermore, a new set of tools is available to judge the goodness of the fit.

- A pairwise **CorMap** analysis in automatically performed between each original $I(t)$ vs. t chromatogram and its reconstruction based on the sum of the fitting Gaussians, for each q- value.
- A new kind of plot becomes available, visualized by pressing ***Global fit by q***. The normalized χ^2 (diamonds connected by a line) and the pairwise CorMap P-values (squares) are plotted as a function of the q-value, as shown below:

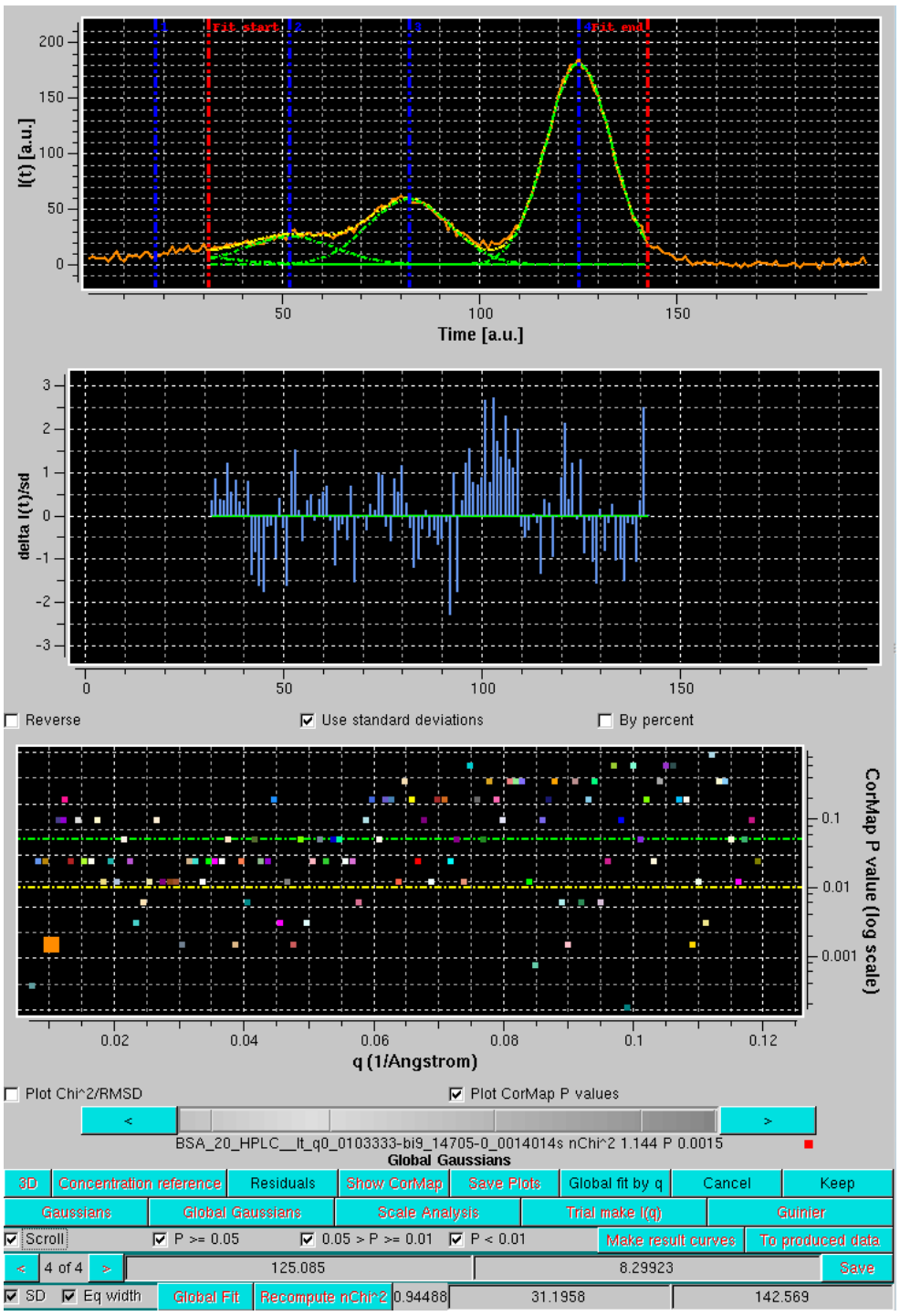


In the *Global fit by q* graph it is possible to visualize either one of or both the two plots, by selecting/deselecting their respective checkboxes positioned just below it (*Plot Chi^2/RMSD* and *Plot CorMap P values*). Note that in the image above, where both plots are shown, their respective *y*-axis scales have been manually modified to allow a better visualization of each plot. The dashed green and yellow horizontal lines mark the usual cut-off *P*-values ($P \geq 0.05$, above the green line; $0.05 > P > 0.01$ between the green and yellow lines; $P < 0.01$, below the yellow line).

Note that the limits of the fit have been moved to exclude the first peak and the tail of the main peak. This was done to concentrate the goodness-of-fit indicators toward the most important part of the fit, including the top $(2/3)^{rds}$ of peaks 2, 3 and 4 (each time the limits are moved, the normalized χ^2 and *P*-values are recomputed by pressing the ***Recompute nChi^2*** button). With these limits, the normalized

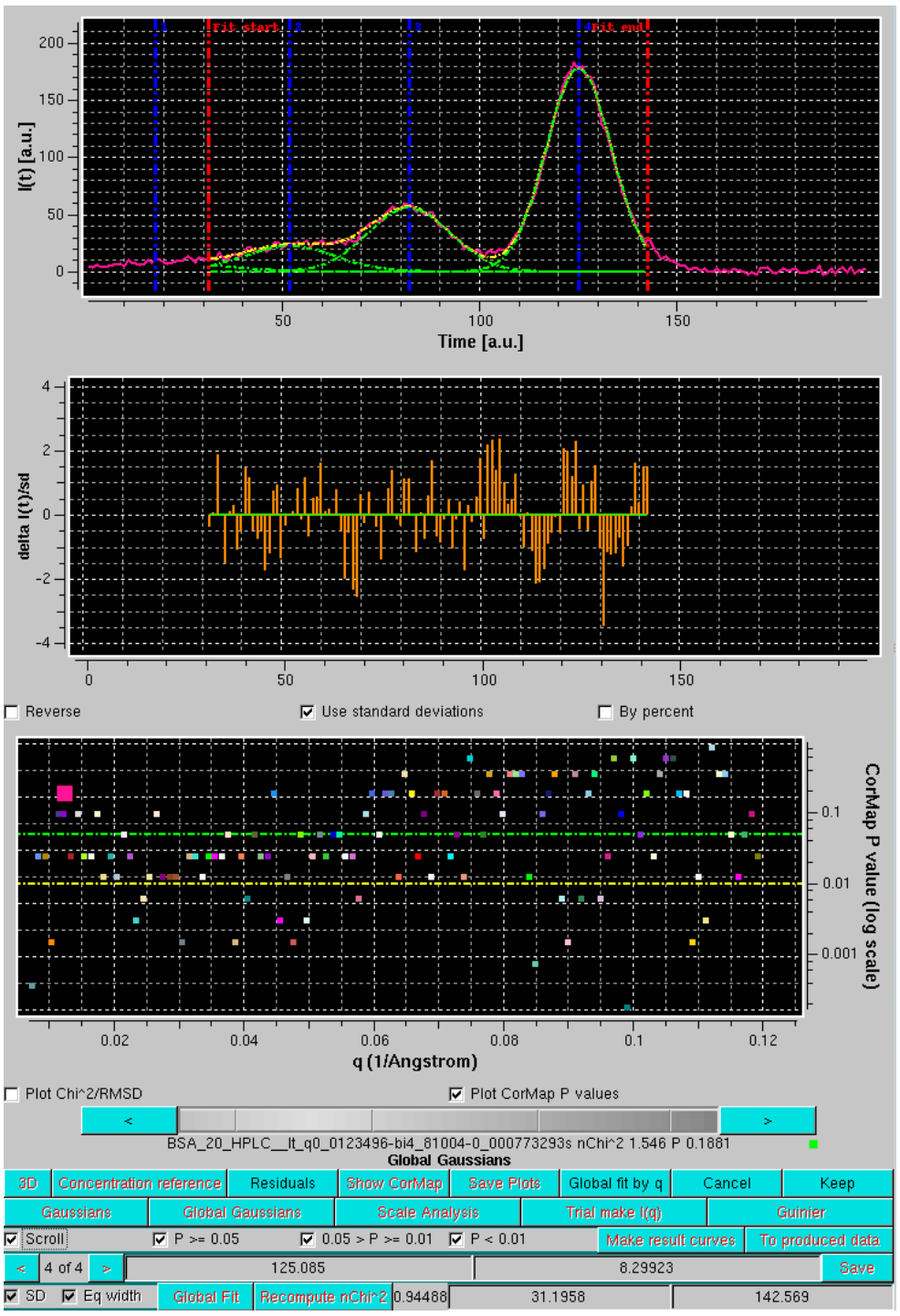
χ^2 values display "reasonable" values, ≈1.5 for the lowest q angles (up to q ≈0.035 Å^{-1}), then almost linearly decaying to ≈0.5 for q ≈0.8 Å^{-1}, being stable afterwards, for a global χ^2 ≈0.95. Likewise, the CorMap *P*-values show a slight trend toward better values as the q increases, but the distribution of really "bad" *P*-values appears to be substantially random.

The correlation between the goodness-of-fit indicators and the distribution of the residuals can be examined for each original/fit *I*(*t* vs. *t* pair by selecting the *Scroll* checkbox:



In the image above, only the CorMap *P*-value are shown. The current chromatograms pair is highlighted in the CorMap *P*-values plot by an enlarged symbol (orange square in this case). Scrolling is performed by either using the grey-scale bar-wheel, or by clicking on the the "<" and ">" buttons placed at its sides. By selecting/deselecting the three checkboxes next to the *Scroll* checkbox (*P >= 0.05*, *0.05 > P >= 0.01*, *P < 0.01*), only the subset(s) whose *P*-values are within those of the selected chechbox(es) will by scrolled.

In the example shown above, by examining the residuals' plot it is clear that the "bad" *P*-value it is due to a poor fit in the inflection point betweern the 3 rd and 4th peaks. If we examine a chromatograms pair just two q values above the first one examined in detail, we can see that oscillations in this zone produce an excellent *P*-value, although this still appears to be a difficult zone to fit with symmetric Gaussians:



The noticeable worse fitting and the end of the main peak could indicate either a slight non-pure Gaussian shape of the peak, or the presence of a small amount of some trailing material in this region.

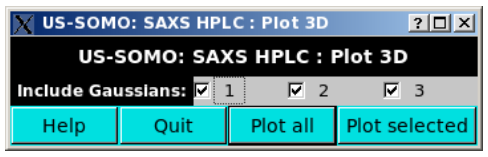
It is best to first ***Save*** and then ***Keep*** the results, and then select all the available *I(t)* vs. *t* chromatograms (use ***Select all*** if only *I(t)* vs. *t* data are present in the **Data files** section).

Global Gaussians can now be applied to all the selected chromatograms. Again, if datasets having points with missing or NaN std. dev. values are found, a pop-up panel will appear (not shown).

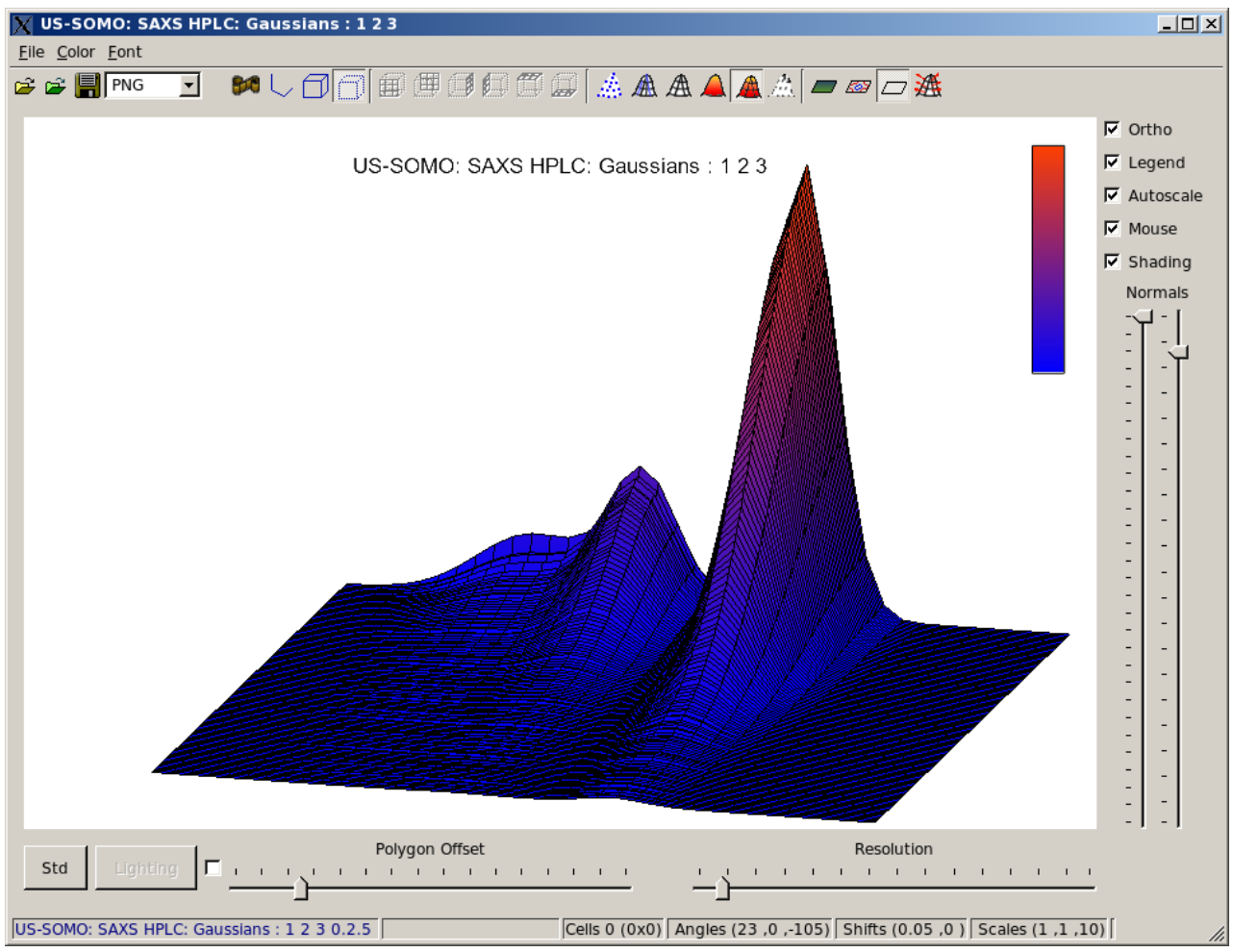
The image above shows the global Gaussians results after applying the global fit parameters found on a subset of data to all chromatograms. Note that in the series of graphs above, the residuals *x*-axis was rescaled (using the graph controls accessed by right-clicking on the graphics window plots) to align it with the selected fit region delimited by the two red vertical lines.

Save and ***Keep*** can then be sequentially pressed to store and accept the global Gaussian results.

Pressing the ***3D*** button will generate a 3D plot of the data, allowing easier detection of potential fitting issues. First a pop-up window will appear:



More infos about this small module can be found here.

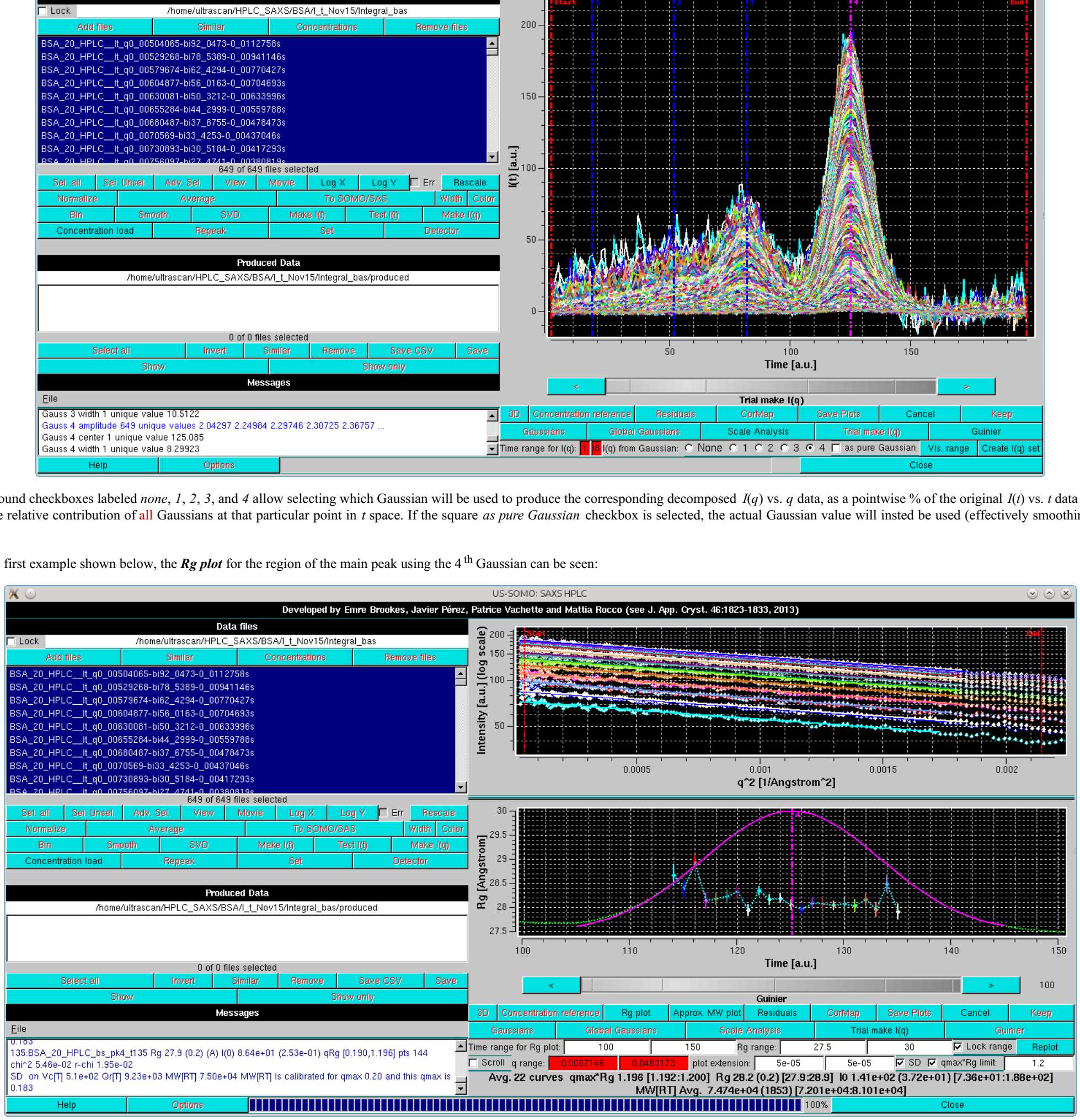


This interactive plot can show any selected set of the Global Gaussians over any collection of curves. The interface is fully interactive for rotations, scaling and zooming along with multiple display and save controls. Its utilization is helpful for visualizing the quality of the global fit.

After Gaussian decomposition, the ***Trial make I(q)*** procedure can be repeated. First, all the available $I(t)$ vs. t chromatograms for which Gaussians have been produced are selected, and the ***Trial make I(q)*** button is pressed. The third commands row under the graphics window will now show additional options:

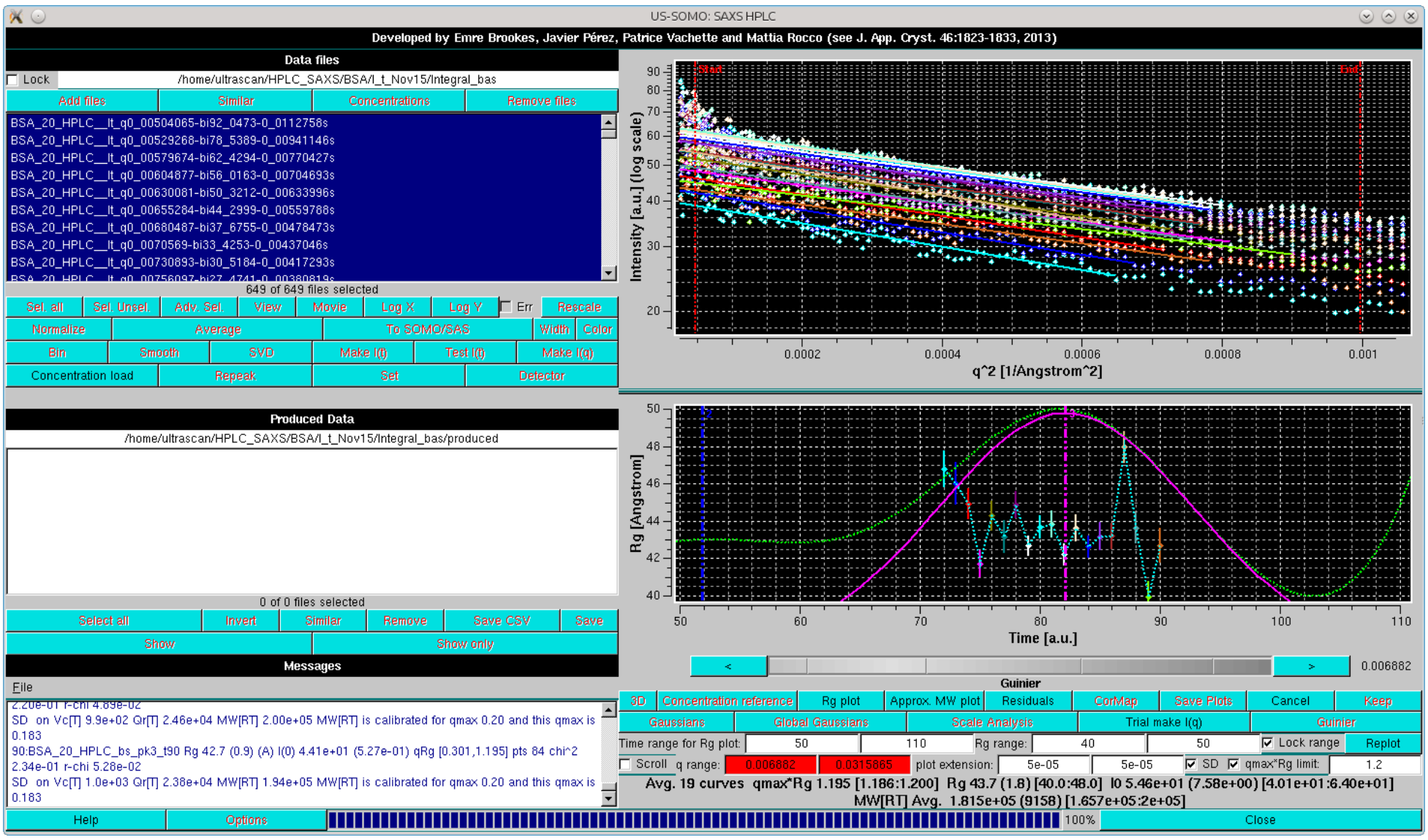
The round checkboxes labeled *none*, *1*, *2*, *3*, and *4* allow selecting which Gaussian will be used to produce the corresponding decomposed $I(q)$ vs. q data, as a pointwise % of the original $I(t)$ vs. t data based on the relative contribution of all Gaussians at that particular point in t space. If the square *as pure Gaussian* checkbox is selected, the actual Gaussian value will insted be used (effectively smoothing the data).

In the first example shown below, the ***Rg plot*** for the region of the main peak using the 4 th Gaussian can be seen:



The contribution of the 3rd Gaussian under the main peak can be now evaluated:

As can be seen, there is a slight contribution (see the ~500-fold reduction in intensity in the Guinier plot) from the peak preceeding the main peak, evidenced by the fairly constant but considerably higher R_g values (≈ 44.5 Å), likely identifying this material as a BSA dimer (also confirmed by the average MW[RT] value).
If we perform this analysis selecting the top region of the dimer peak (frames 72-90), we find similar R_g values, ≈ 43.7 Å, almost evenly distributed across the peak:



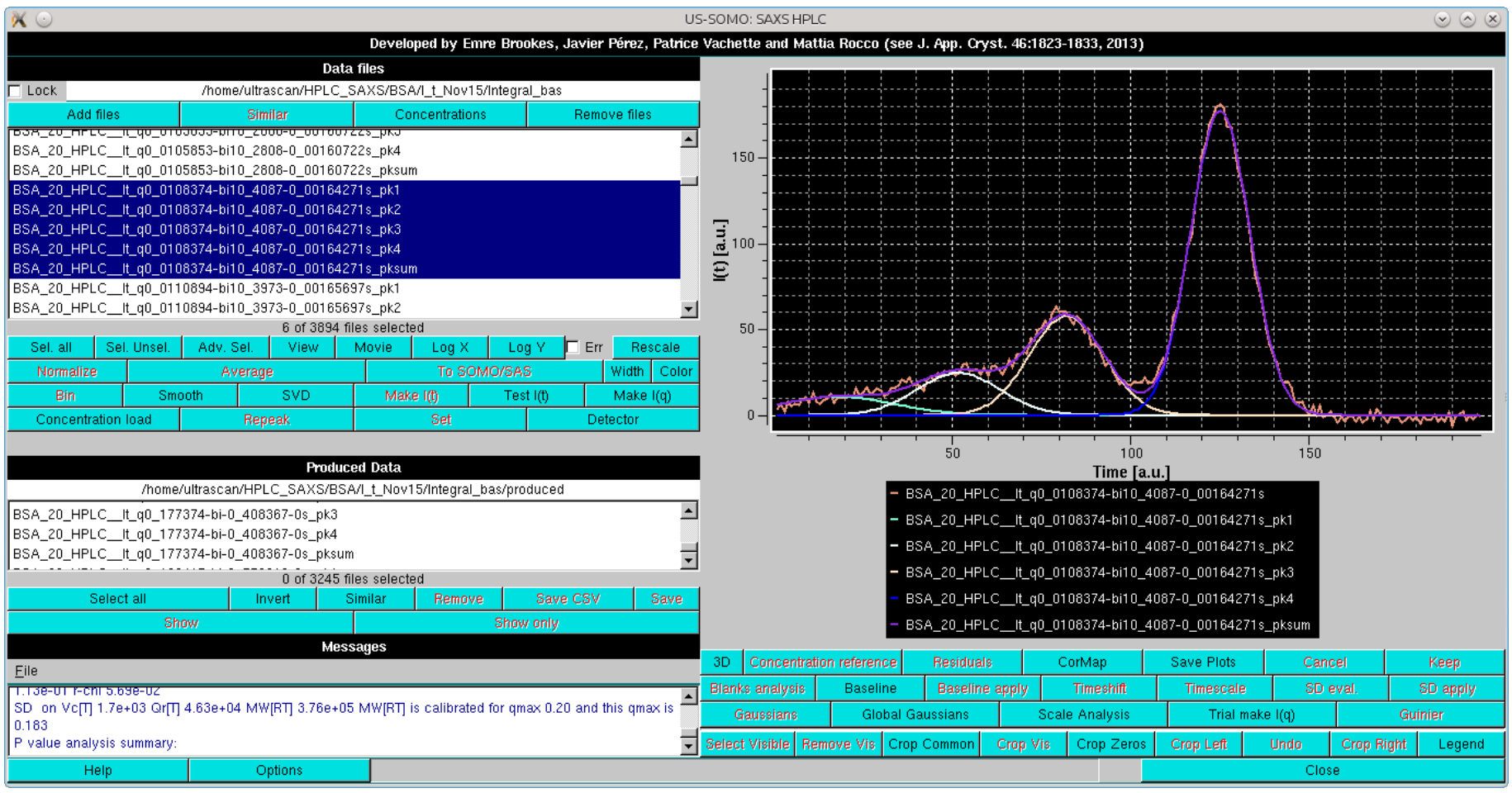
Note that some of the curves present an upward curvature at low q-values, likely indicating a non-ideal baseline correction for this sample.
Finally, the contribution of the 2nd Gaussian under the trimer peak is evaluated:

finding higher R_g values (≈ 60 Å) and a reasonably flat distribution.

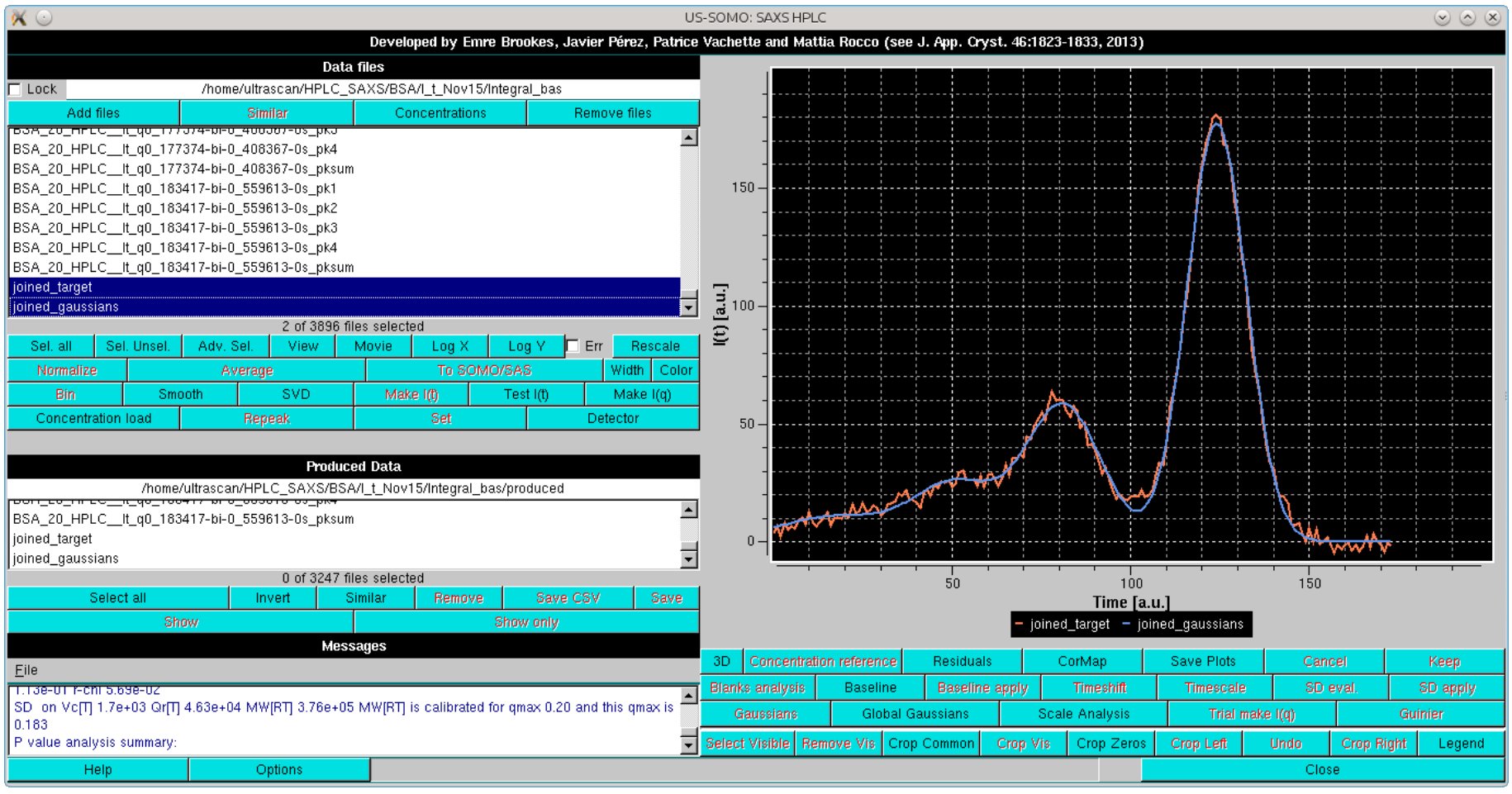
The ***Scroll*** feature is also available, allowing the examination of the individual Guinier plots. Likewise, the ***Approx. MW plot*** can show the approximate molecular weight values calculated with the Rambo and Tainer method (Accurate assessment of mass, models and resolution by small-angle scattering. Nature 496:477-481, 2013) . However, as it requires to produce meaningful results a more extended q range than that available for the HPLC-SAXS BSA study presented here, we will not present such plots here.

Each individual Gaussian is defined by three numbers: the amplitude, width and center. As such, they are not "curves" in the sense of the loaded files, which are collections of data points. Therefore, the Gaussians can not be visualized with the facilities of the program outside of Gaussian or Global Gaussian modes.

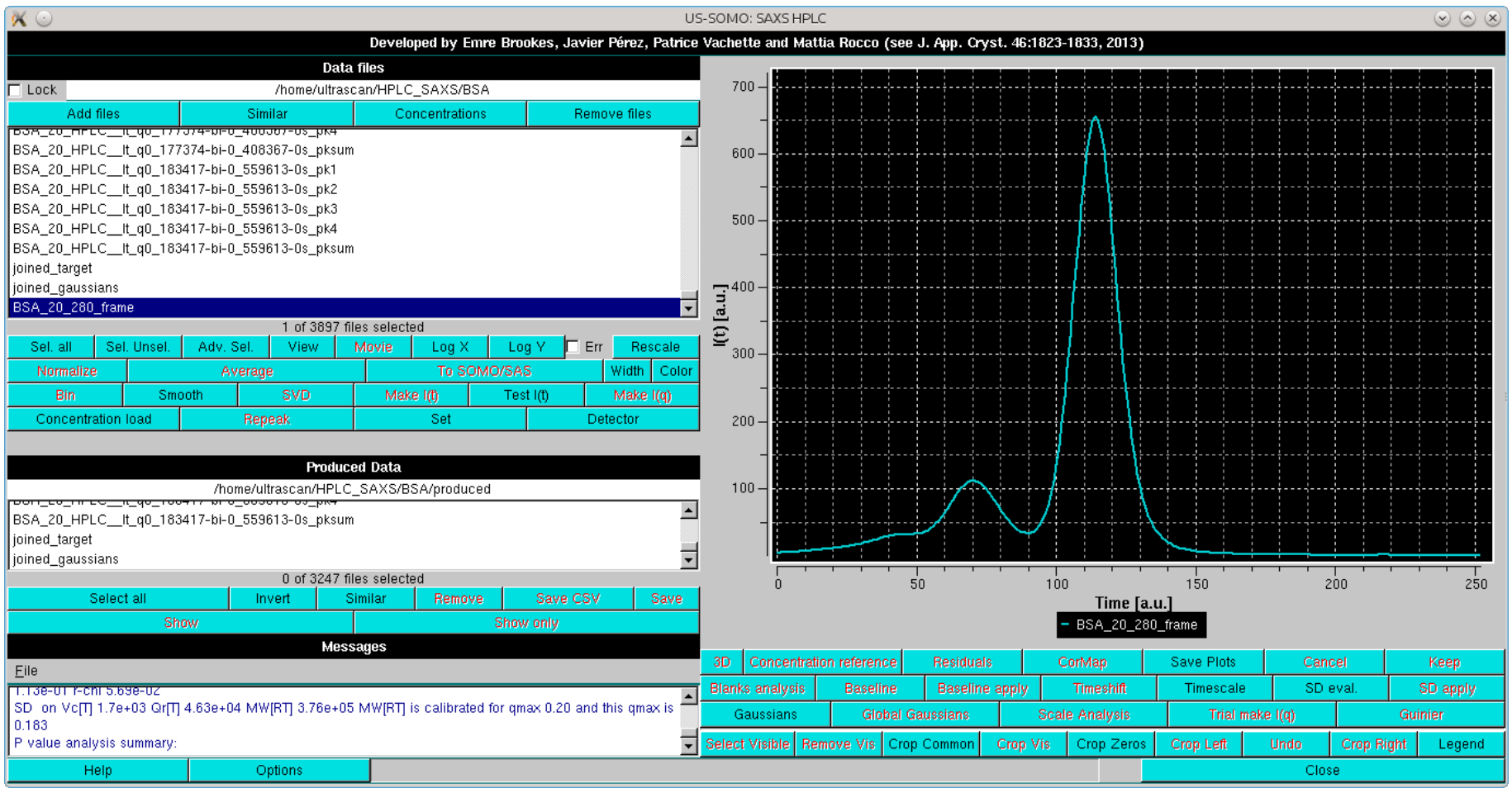
To allow the visualization of the Gaussians, the ***To produced data*** button is provided which produces curves of individual Gaussians and their sum. This is available in either Gaussian or Global Gaussian modes. The resulting curves are collections of data points that can be visualized outside of the Gaussian modes. The Global Fit method requires a simultaneous fit of all the selected curves. This is internally represented by joining all the selected curves along the time/frame dimension to produce one long curve. Of course, each curve is generally on the same time/frame axis range, so to maintain increasing time/frame numbers, curves subsequent to the first one are placed into the joined curve with an offset in time/frame.



To visualize the joined curve and the Global Gaussian fit to the joined curve the ***Make result curves*** button is provided. This will create the joined curve along with the joined Global Gaussian fit as a pair of curves that can be visualized outside of the Global Gaussian mode, as in the example shown below:

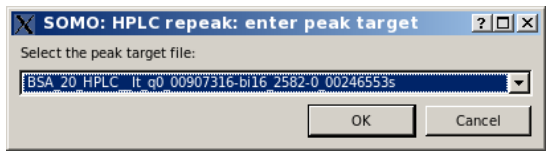


If available, a concentration chromatogram deriving from UV or refractive index monitors can now be processed. After uploading a suitable file with the ***Concentration load*** button:

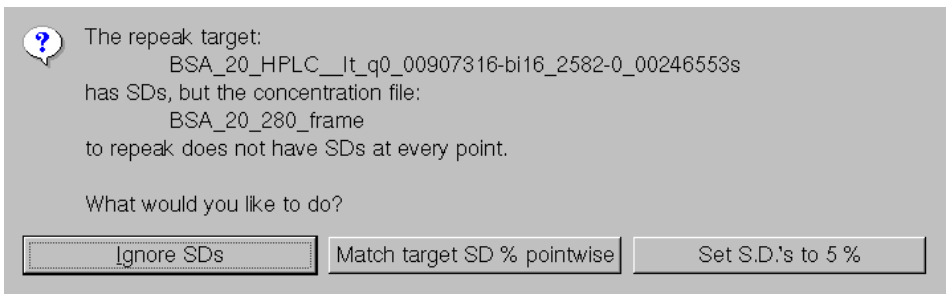


the first operation is to rescale it to one of the high intensity but relatively low-noise *I(t)* vs. *t* chromatograms. This is done by selecting the two files:

and pressing ***Repeak*** in the left-side command panel, which will bring up a small window asking to identify the target chromatogram (in case multiple were selected).



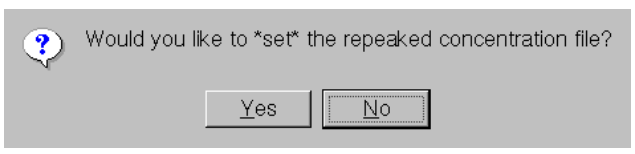
Usually concentration detector data have no associated SDs. In this case, another pop-up panel will appear presenting three options:



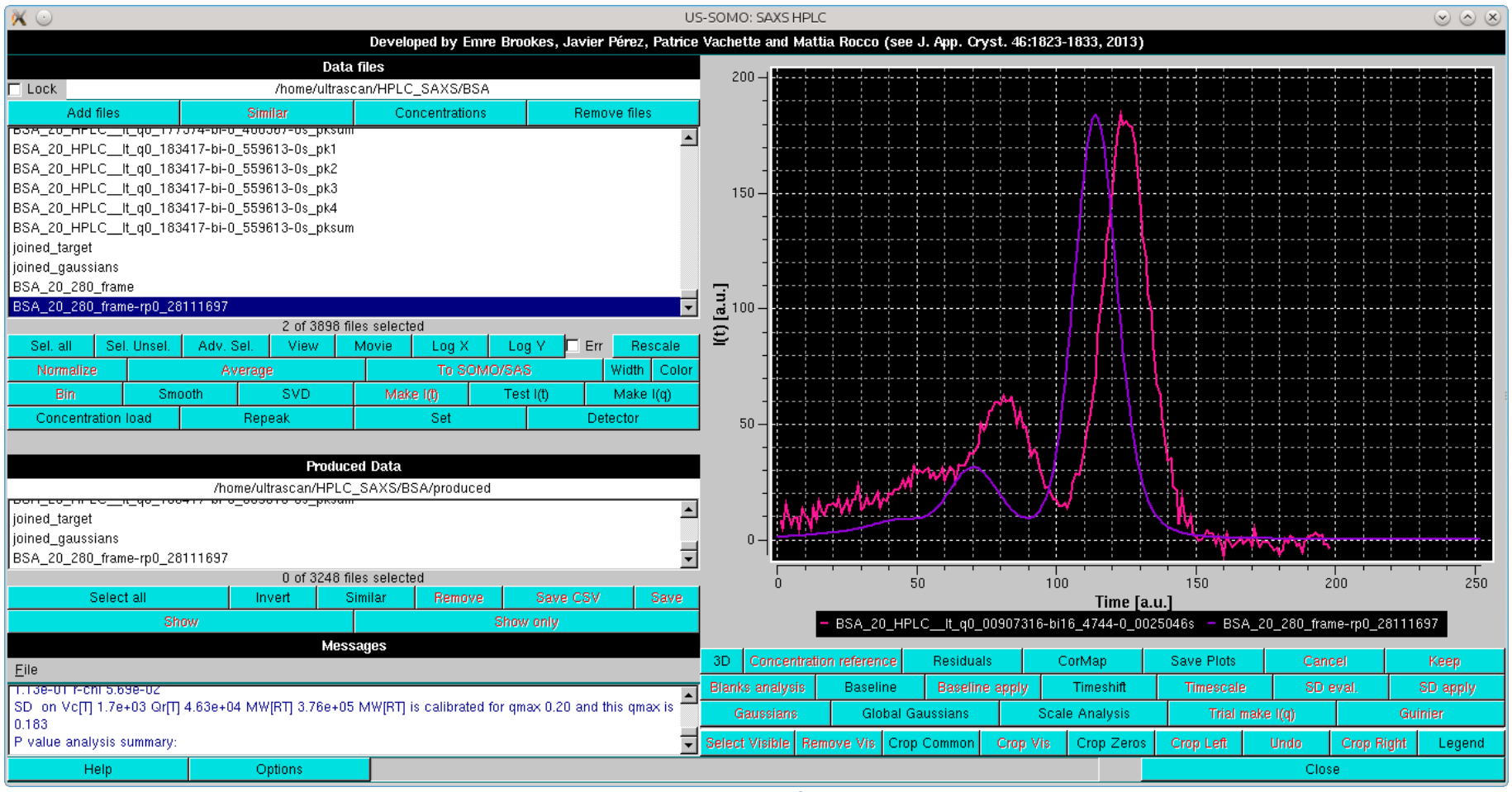
- *Ignore S.D.s*.
- *Match target S.D. % pointwise*; this will generate artificial SDs data associated with the concentration detector data matching %wise and pointwise the SDs present in the target *I(t)* vs. *t* data.
- *Set S.D.s to 5%*; this will generate artificial SDs data at a 5% level associated with the concentration detector data.

Selecting the second option will allow a consistent Gaussian analysis of the concentration detector data.

The repeak operation is then automatically performed, and another pop-up message will appear:

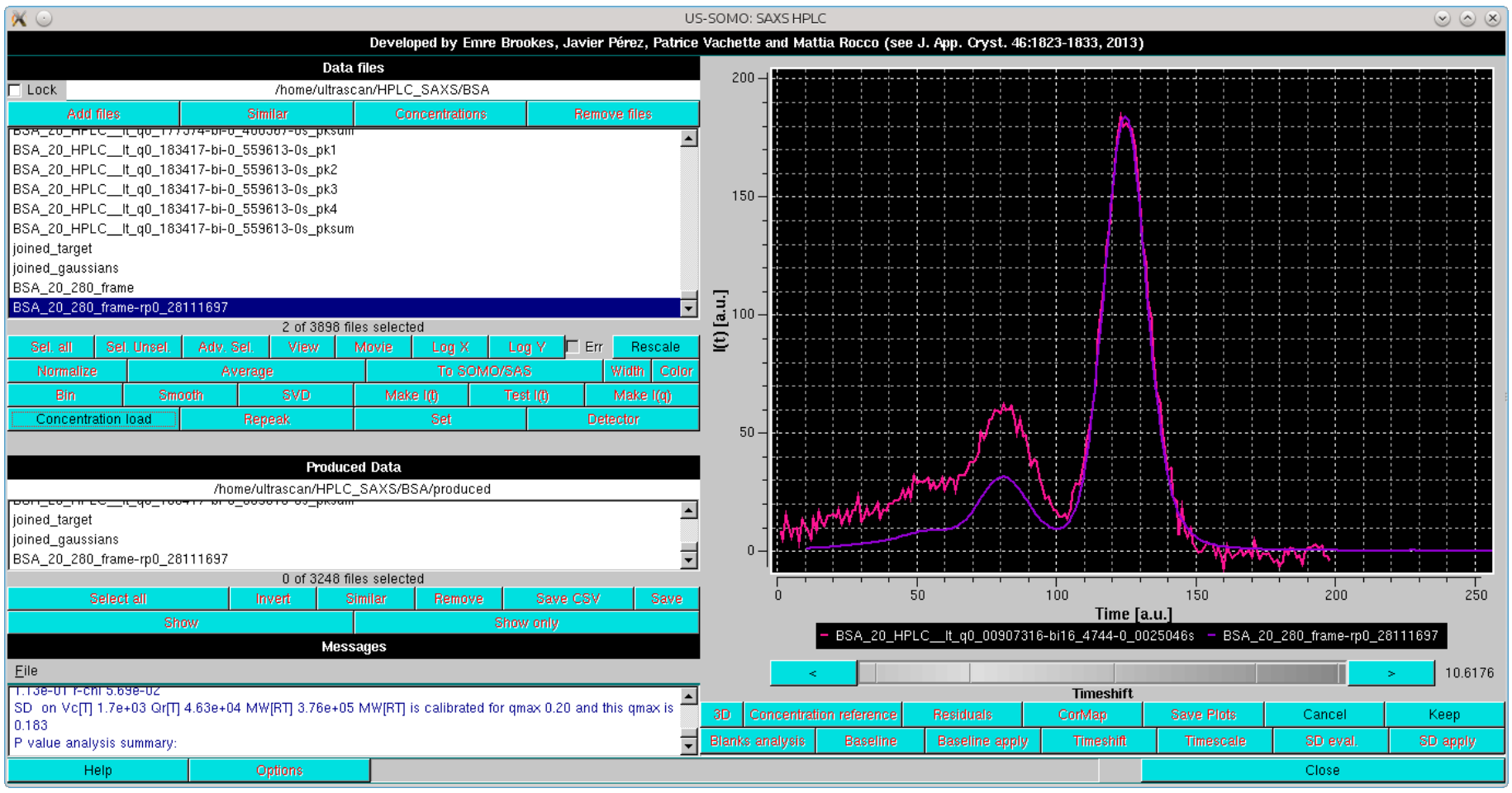


If no time-shift between the concentration and the SAXS detectors is present, the repeaked concentration file can then be directly associated with the SAXS datasets. Otherwise, it is better to perform first the timeshift operation (see below) before setting the concentration file. The result of a re-peak operation is shown below, and the scaling factor is added to the concentration dataset filename.



After re-peaking, the concentration chromatogram usually must be time-shifted to align its peaks to the *I(t)* vs. *t* chromatograms using the ***Timeshift*** button.

Again, at least two files must be selected, one is the concentration data, the other belonging to the *I(t)* vs. *t* (the file used for re-peaking is normally used for this operation). An automatic alignment is first performed, using the highest intensity peaks. The alignment can then be refined manually by left-clicking and moving the mouse over the grey-shades wheel bar below the graphics window until the two chromatograms are best aligned.

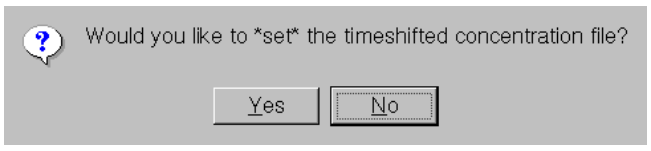


The value of the timeshift is reported in the field next to the grey-shades wheel bar.

Cancel will stop the operation.

Keep will keep the time-shifted data. The produced data will have the timeshift value added to its filename on saving.

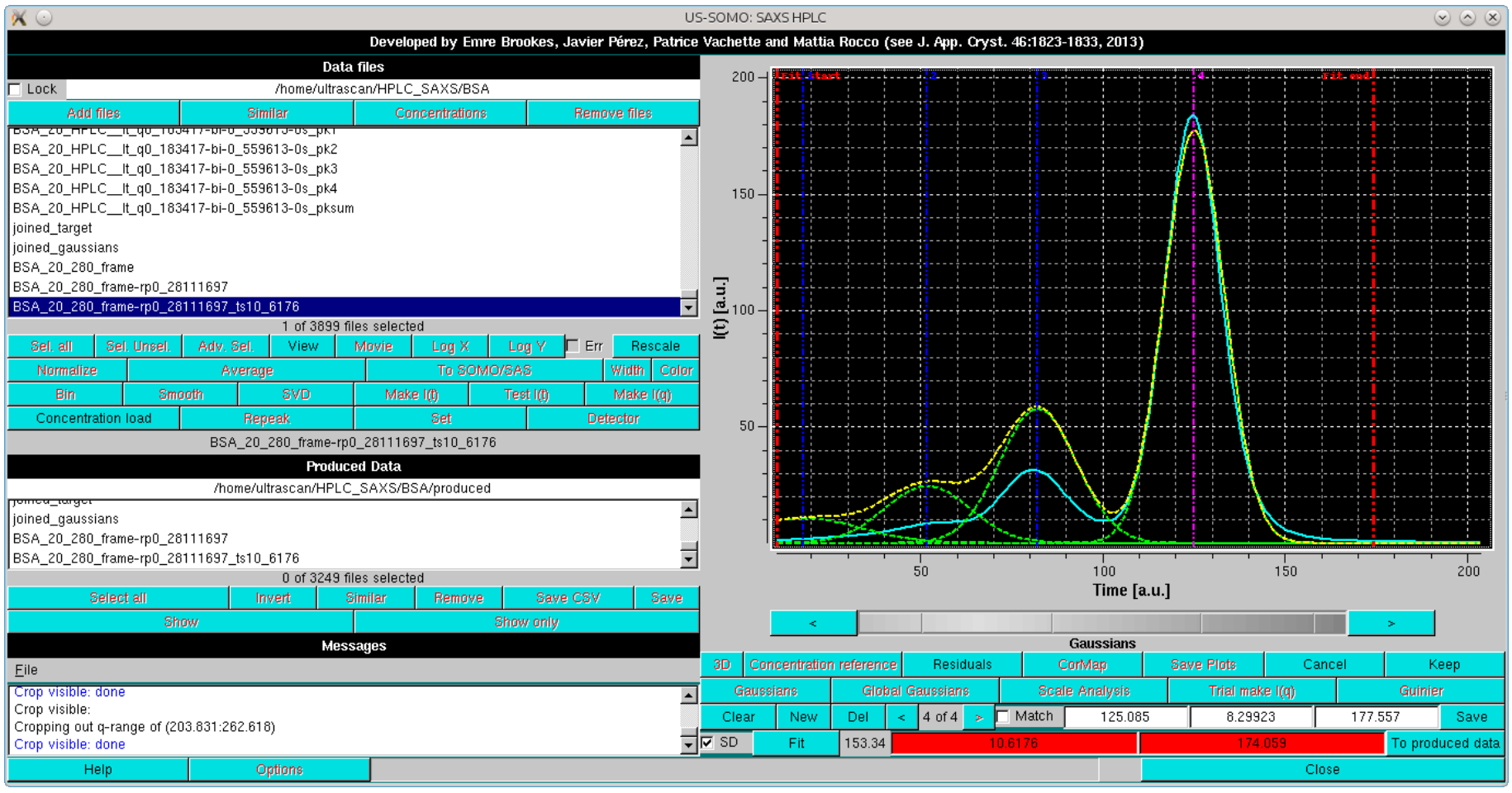
Another pop-up panel will appear, asking to associated the time-shifted concentration file to the SAXS data:



Answering "yes" will then associate the re-peaked, time-shifted concentration data to the *I(t)* vs. *t* SAXS dataset under analysis. This operation can be anyway performed at any time by selecting only a concentration chromatogram dataset, and pressing ***Set***. The concentration chromatogram shown in this example is then cropped to have approximately the same frame number as the SAXS *I*(*t*) vs. *t* chromatograms.

The re-peaked, time-shifted concentration chromatogram can be now fitted with Gaussians, using for initialization the set derived from the *I(t)* vs. *t* chromatograms (note: it is mandatory that the same number of Gaussians be used for both the concentration and *I(t)* vs. *t* chromatograms).

This is done by first selecting only the concentration chromatogram and then pressing ***Gaussian***, which will bring up the current Gaussian parameters automatically rescaled to the highest intensity in the concentration chromatogram:

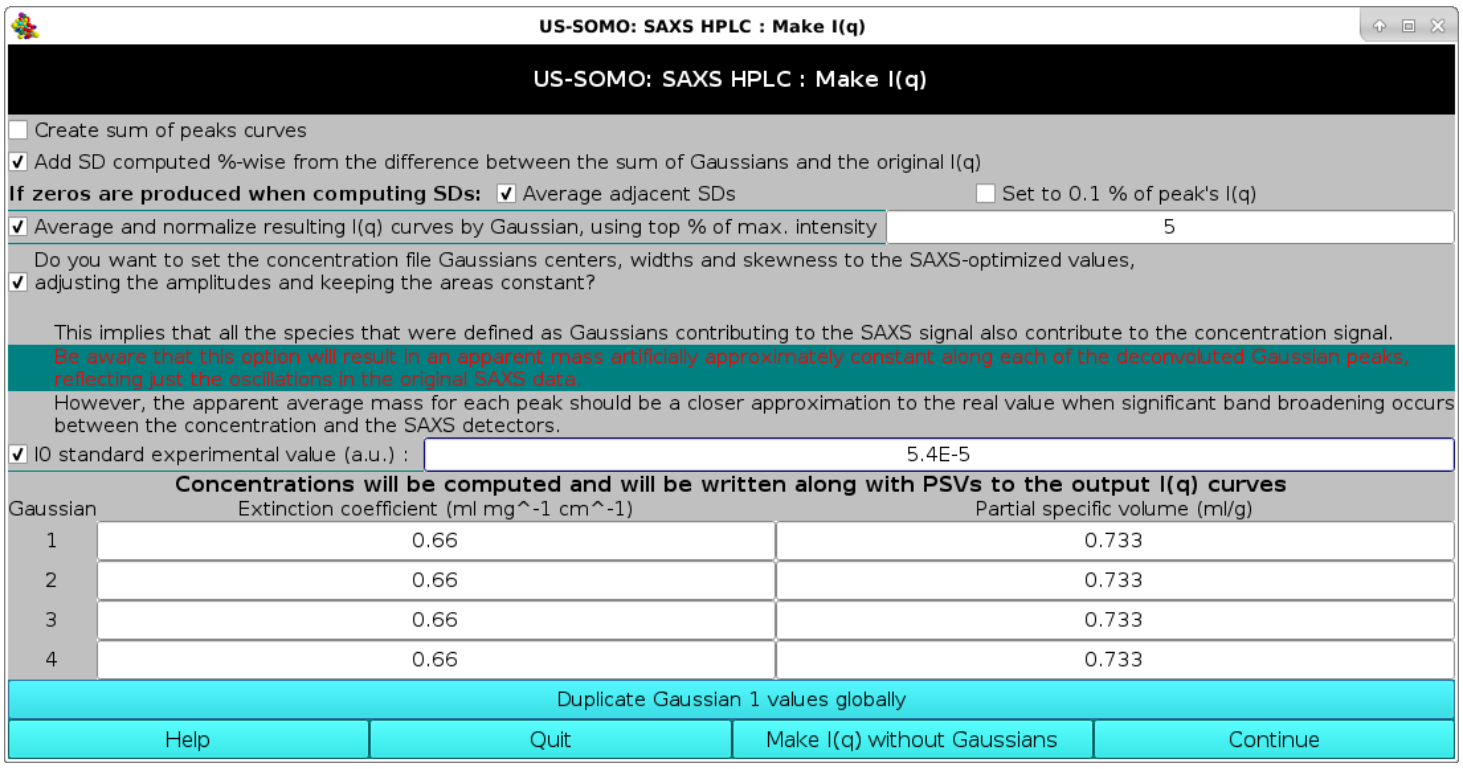


Pressing ***Fit*** will then bring up the **Fit** window (see here) and an initial round is done by keeping fixed both the position and widths. If necessary, a refinement can be done by keeping fixed the smallest, front eluting peak(s), and allowing only a limited shift to a % of the initial values from the widths and positions determined from the SAXS data (suggested: 2-3% max). This should compensate for slight misalignment between the concentration and SAXS detectors chromatograms.

As evidenced in the image above, some band broadening has occurred between the UV-VIS and SAXS detectors. While the issue appears to be relatively minor here, it can be more serious. To at least partially mitigate this issue, we have implemented a re-shaping routine that re-aligns the shape of the concentration detector chromatogram to that of the SAXS detector chromatograms. It is based on determining first the area under each Gaussian peak in the concentration chromatogram after fitting it with the SAXS-derived Gaussians with minimal centers and widths changes, as described above. Then, when the ***Make I(q)*** routine is launched, the concentration chromatogram Gaussians can be optionally re-shaped on the SAXS-optimized Gaussians, keeping their areas fixed and adjusting the other parameters (see below).

The ***Save*** and ***Keep*** buttons must be then pressed to store and associate the resulting Gaussians to the concentration chromatogram. On re-generating the *I(q)* vs. *q* frames (see below), each concentration Gaussian peak will be mapped onto the corresponding *I(t)* vs. *t* peaks.

The ***Make I(q)*** button becomes available every time that more than one *I*(*t*) chromatogram is selected. If Gaussian fitting was performed, pressing it will produce a series of *I(q)* vs. *q* curves for each Gaussian peak for each frame of the chromatogram on which the global operations have been carried out. An option panel in a pop-up window will allow several choices:

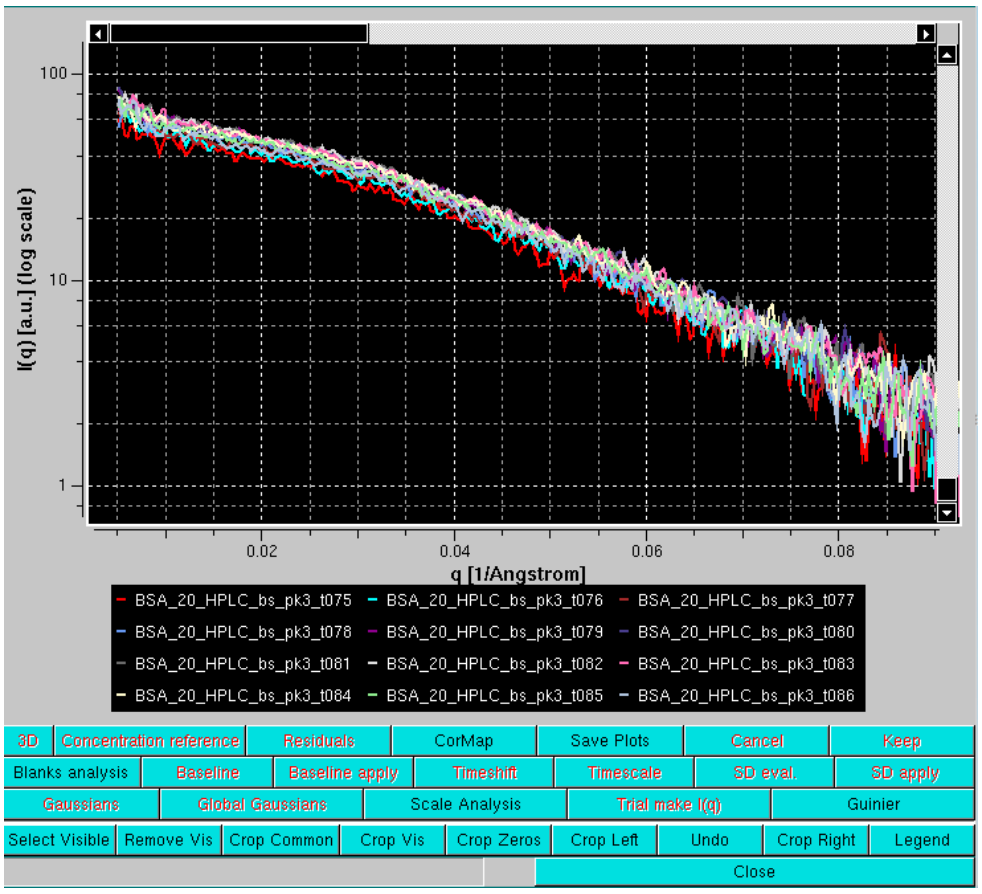


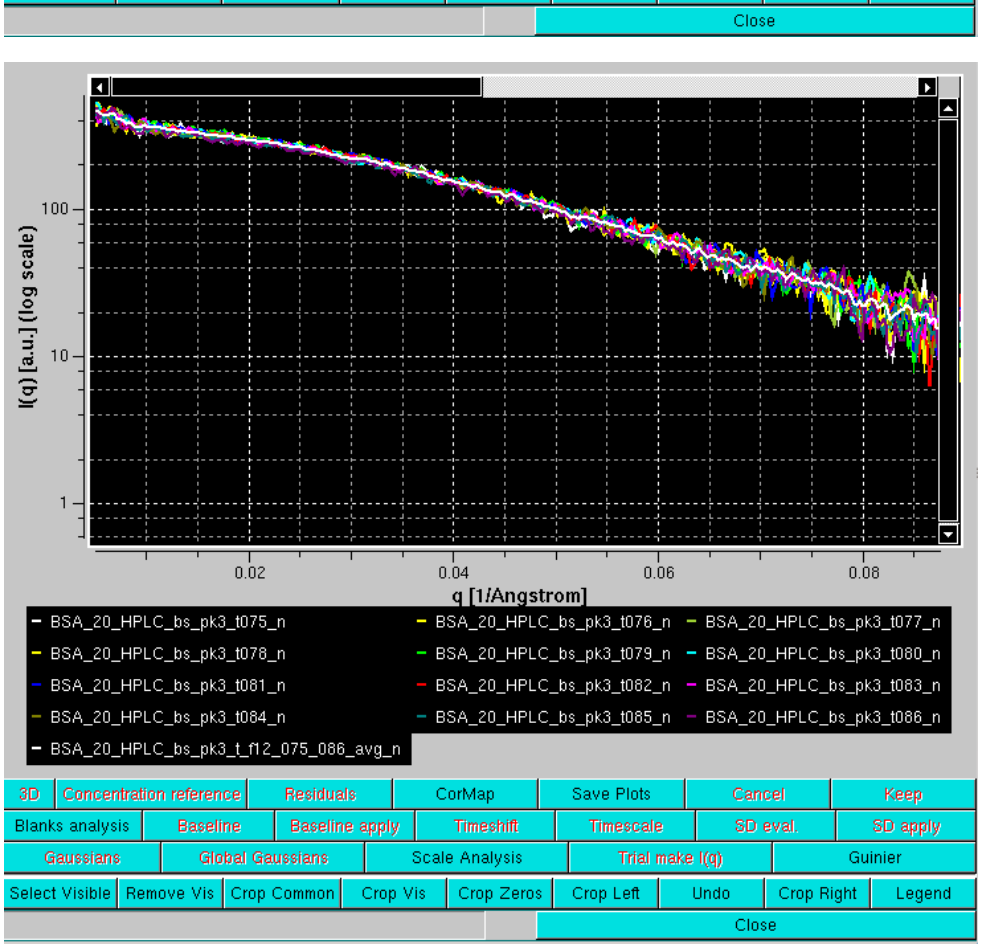
A description of this module can be found here.

Once the *I(q)* vs. *q* files have been generated, it is possible to view the resulting Gaussian contributions and their sum, as shown below:

Here the original and decomposed *I(q)* for frame #59 (cream), originally presenting an overlap between the trimers and dimers peaks, are shown, together with the reconstructed sum with baseline back-addition (see the legend for details). The drop of intensity at q values > 0.07 Å^{-1} for Gaussian peak #1 (purple) is due to its contribution vanishing in the high q range. Note how there is a contribution from Gaussians #1, #2 (orange), and #3 (olive green) in this frame (peak #4, not contributing at all in this frame, is not shown). Note also how the reconstructed curve with baseline back-addition (whose color would be cyan) perfectly superimposes with the original frame data (cream), and thus is almost not visible in this frame except at the very end of the q range.

A zoom into the low q region for frames #75-86 of the Gaussian peak #3 it is shown in the next two images, before and after concentration normalization (***Normalize*** button):



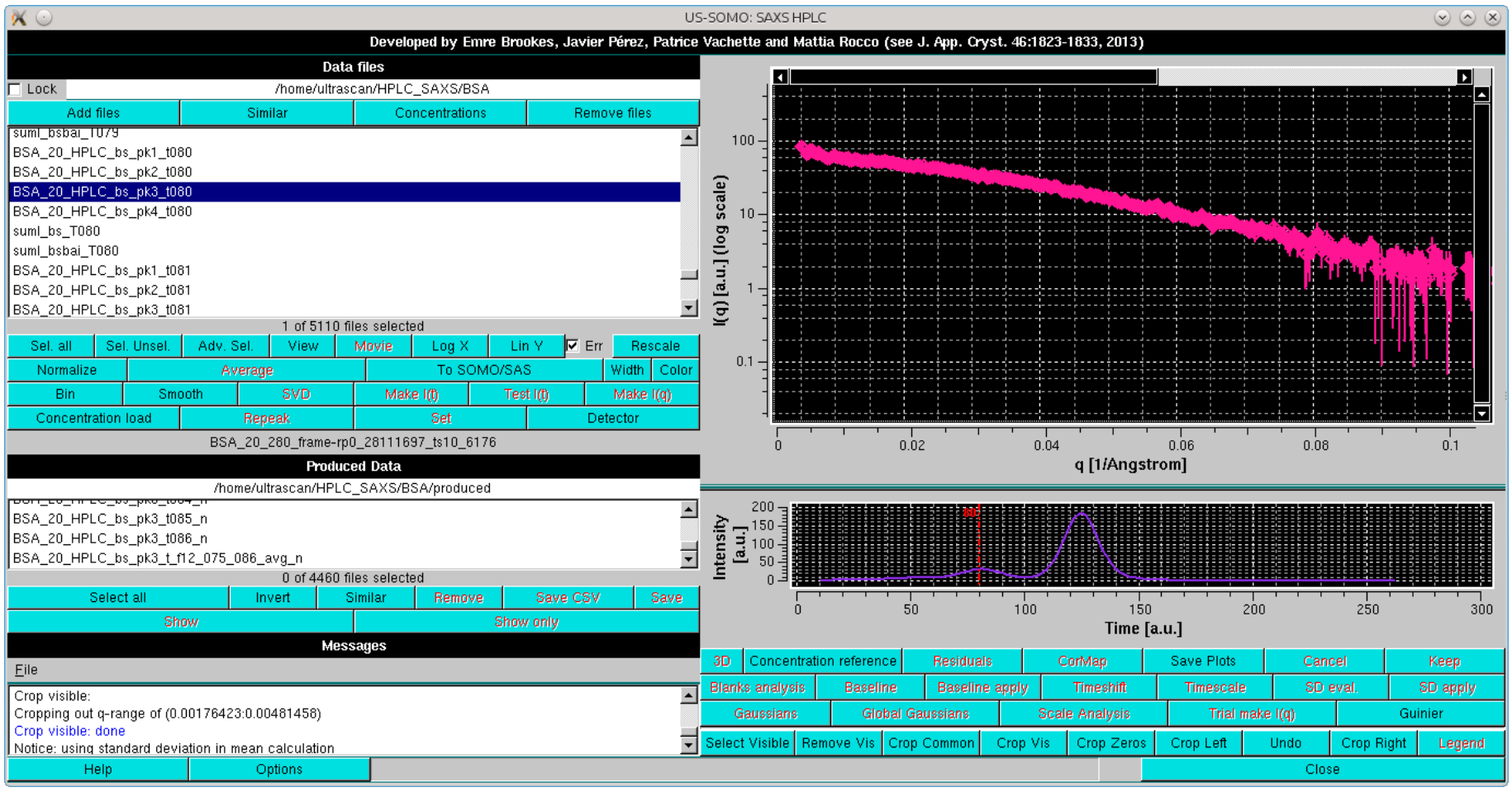


In the latter, an average curve (obtained by pressing the ***Average*** button) is also superimposed. Such re-generated *I(q)* vs. q data can be directly exported in the main **US-SOMO *SAS*** module for further

operations by pressing the ***To SOMO/SAS*** button.

Note that starting from the January 2018 release, an automatic selector of frames to be averaged has been introduced within the ***Make I(q)*** option panel (see here). This option will also normalize each frame by its associated concentration value, if present, before performing the average.

If a concentration chromatogram is associated with the data, an additional utility present in this module allows to map a single selected *I(q)* vs. *q* dataset onto the concentration chromatogram, by pressing the ***Concentration reference*** button:



In the example shown above, the *I(q)* vs. *q* data for the decomposed peak #3 frame #80 are shown with their associated errors, and below it the position of this dataset is shown by the vertical red line on the associated concentration cromatogram. Each time a different chromatogram is selected, its position will be mapped on the concentration plot.
Pressing the ***Concentration reference*** button again will make this additional plot disappear.

Finally, the data shown in any of the plots currently visualized can be saved in csv-formatted files by pressing the ***Save plots*** button. This will open a pop-up dialogue window where the location and the root filename for the cvs files can be set.

---

www contact: Emre Brookes

SOMO HPLC-SAXS Module Gaussian analysis theory

Last updated: December 2017

NOTICE: the Gaussian decomposition method is being developed by E. Brookes, J. Perez, P. Vachette, and M. Rocco.
Portions of this help file are taken from the Supplementary Materials of Brookes et al., "Fibrinogen species as resolved by HPLC-SAXS data processing within the UltraScan SOlution MOdeler (US-SOMO) enhanced SAS module", J. Appl. Cryst. 46:1823-1833 (2013), and from Brookes et al. "US-SOMO HPLC-SAXS Module: Dealing with Capillary Fouling, and Extraction of Pure Component Patterns from Poorly Resolved SEC-SAXS Data", J. Appl. Cryst. 49:1827-1841 (2016).

Given the matrix **I** containing columns *i* of *I(q)* and rows *j* of *I(t)*, the principles of Gaussian analysis can be schematized as follows.

Single curve fitting:

Pick a row *i* of **I** and define a set of *p* Gaussians $\left\{G_k^i(t)\right\}_{k=1}^{p}$, with amplitudes a_k^i, centers b_k^i, and widths c_k^i. Then:

$$G_k^i(t) = a_k^i e^{-\frac{(t-b_k^i)^2}{2(c_k^i)^2}}$$

In the **US-SOMO HPLC-SAXS** module, we let the user visually place the centers b_k^i, and subsequently provide several methods for fitting (see below) by minimizing over, in general, 3*p* variables, a_k^i, b_k^i, and c_k^i:

$$\sum_j \left[\left(\sum_k G_k^i(t_j)\right) - I_{ij}\right]^2$$

or in the case that $\forall j: S_{ij} \neq 0$ (i.e., the i^{th} row of the matrix *S* containing the data-associated SDs has no zero elements):

$$\sum_j \frac{\left[\left(\sum_k G_k^i(t_j)\right) - I_{ij}\right]^2}{S_{ij}}$$

In the program, there are options to fix a combination of individual Gaussian curves *k*, amplitudes *a*, centres *b*, and widths *c*, which would result in fewer than 3*p* variables during the minimization. Constraints, in percentage from previous value or from the initial value, are also available for *a*, *b*, and *c*.

Global Gaussians:

In the US-SOMO program, entering the ***Global Gaussian*** mode does a fit of the preset single curve $\left\{G_k^i(t)\right\}$ against every curve *i* = 1, ..., *m*, keeping the centers *b* and widths *c* fixed. This provides an initialization of the amplitudes *a* for all curves as a starting point for global fitting or for refinement/extension to other datasets a previous global fitting on a subset of data.

Global fitting:

Given a $\left\{G_k^l(t)\right\}$ for a specific row *i* = *l* from the result of a single curve fitting, one can globally fit over the amplitudes a_k^i by utilizing common centers, $b_k^i = b_k^l$ for = {1, ...,*m*; $i \neq l$}, and common widths, $c_k^i = c_k^l$ for = {1, ...,*m*; $i \neq l$}, and then doing a global minimization over the *pm* + 2*p* variables a_k^i, b_k^l, c_k^l, as above. Global fitting is currently only available with a Levenberg-Marquardt minimization routine. As in the single Gaussian fitting, there are options to fix a combination of individual Gaussian curves *k*, amplitudes *a*, centres *b*, and widths *c*, which would result in fewer variables during the minimization. Constraints, in percentage from previous value or from the initial value, are also available for *a*, *b*, and *c*.

---

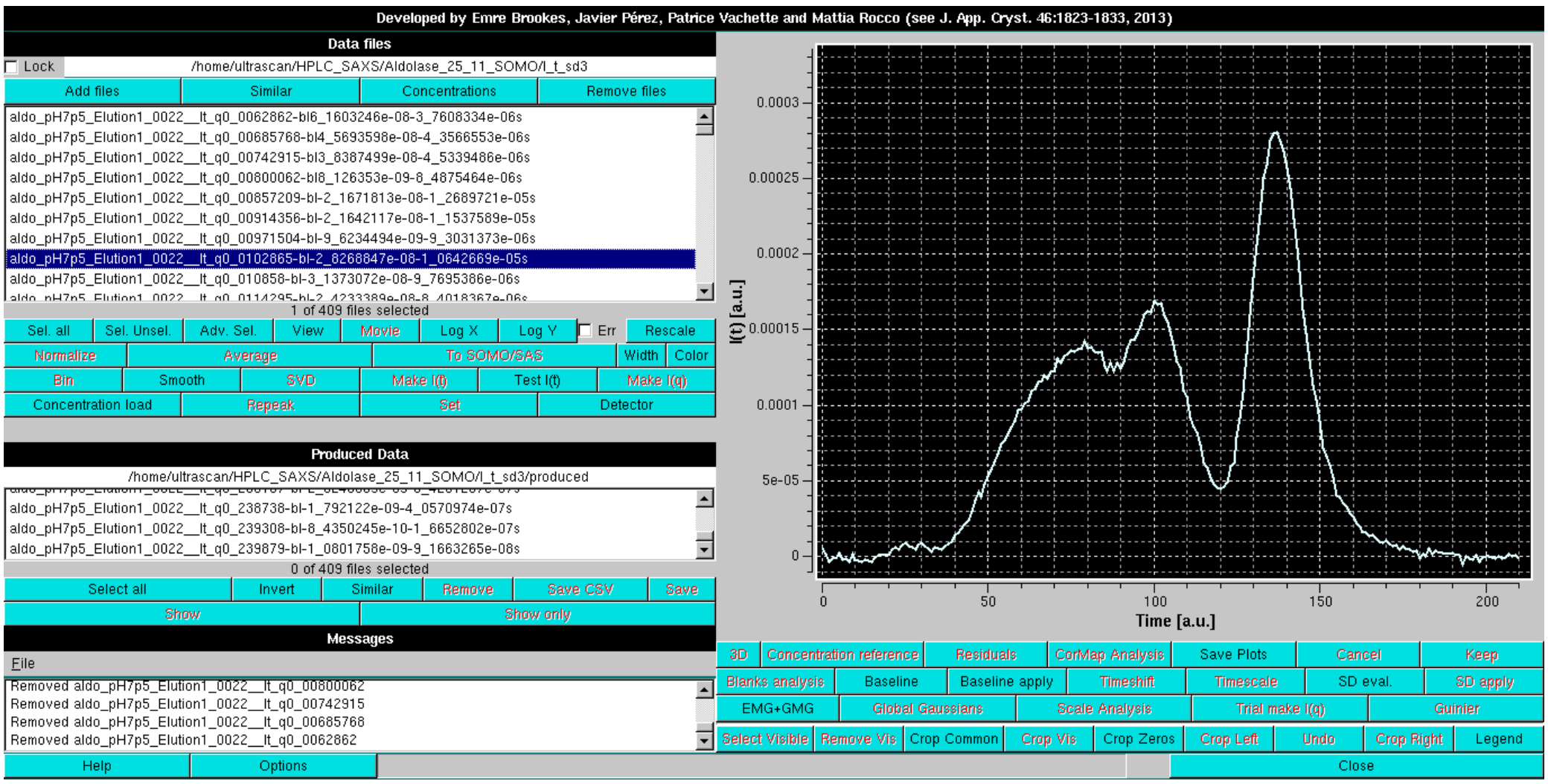
www contact: Emre Brookes

SOMO HPLC-SAXS Module: Gaussians with distorsion(s) operations

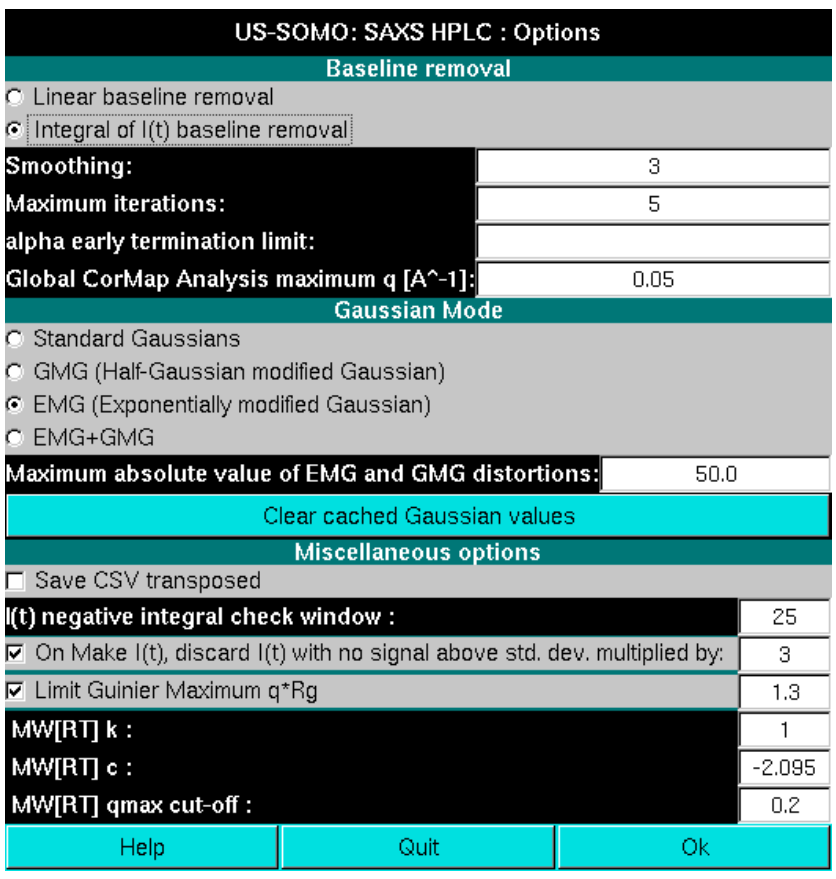
Last updated: December 2017

NOTICE: the Gaussian decomposition method is being developed by E. Brookes, J. Perez, P. Vachette, and M. Rocco.
Portions of this help file are taken from the Supplementary Materials of Brookes et al., "Fibrinogen species as resolved by HPLC-SAXS data processing within the UltraScan SOlution MOdeler (US-SOMO) enhanced SAS module", J. Appl. Cryst. 46:1823-1833 (2013), and from Brookes et al. "US-SOMO HPLC-SAXS Module: Dealing with Capillary Fouling, and Extraction of Pure Component Patterns from Poorly Resolved SEC-SAXS Data", J. Appl. Cryst. 49:1827-1841 (2016).

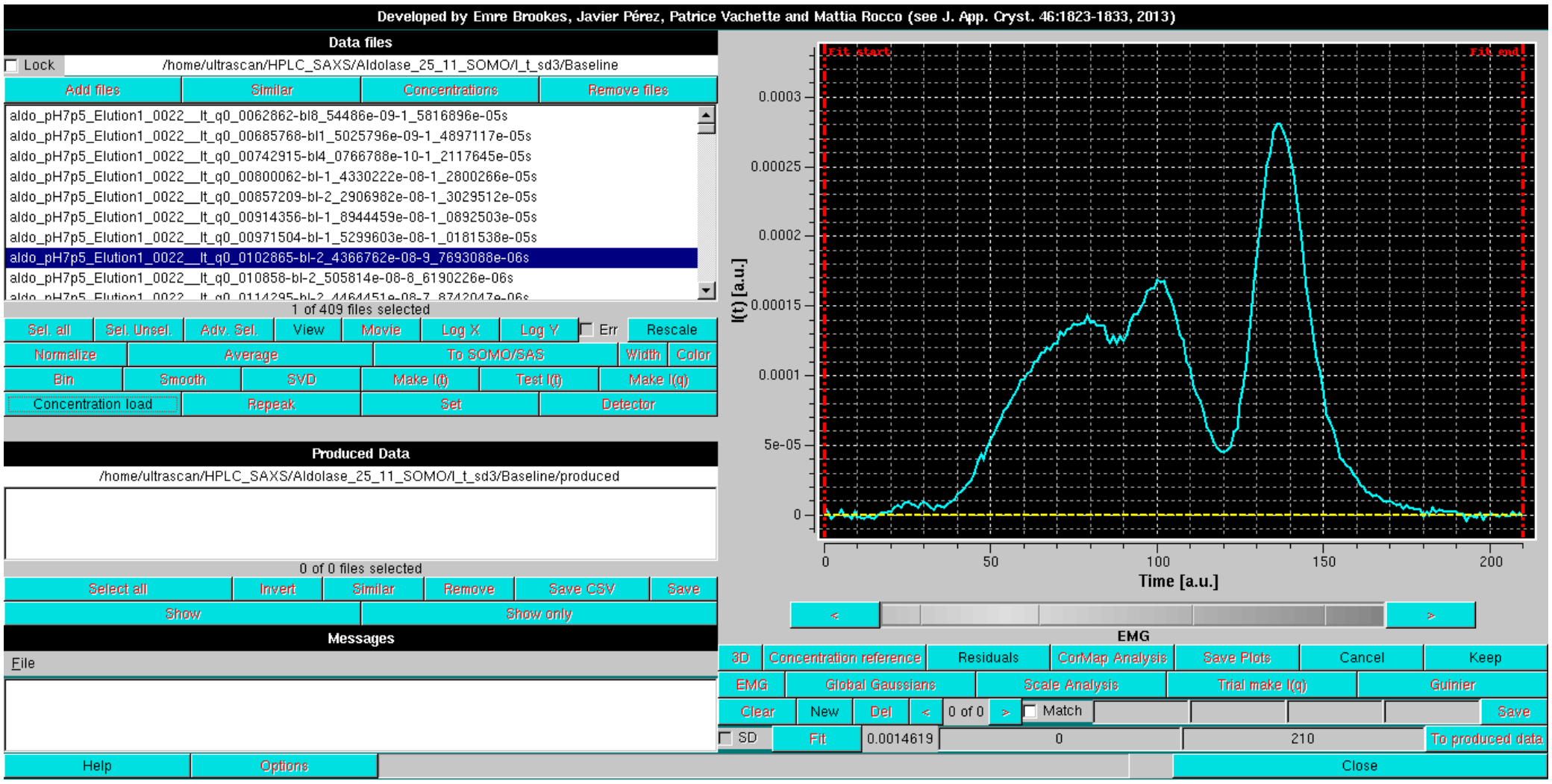
This manual part describes the Gaussian analysis and decomposition operations using non symmetrical (skewed, distorted) Gaussians. The dataset we will use as an example is the same HPLC-SAXS analysis of Aldolase that is used to demonstrate the linear baseline operations (see here):



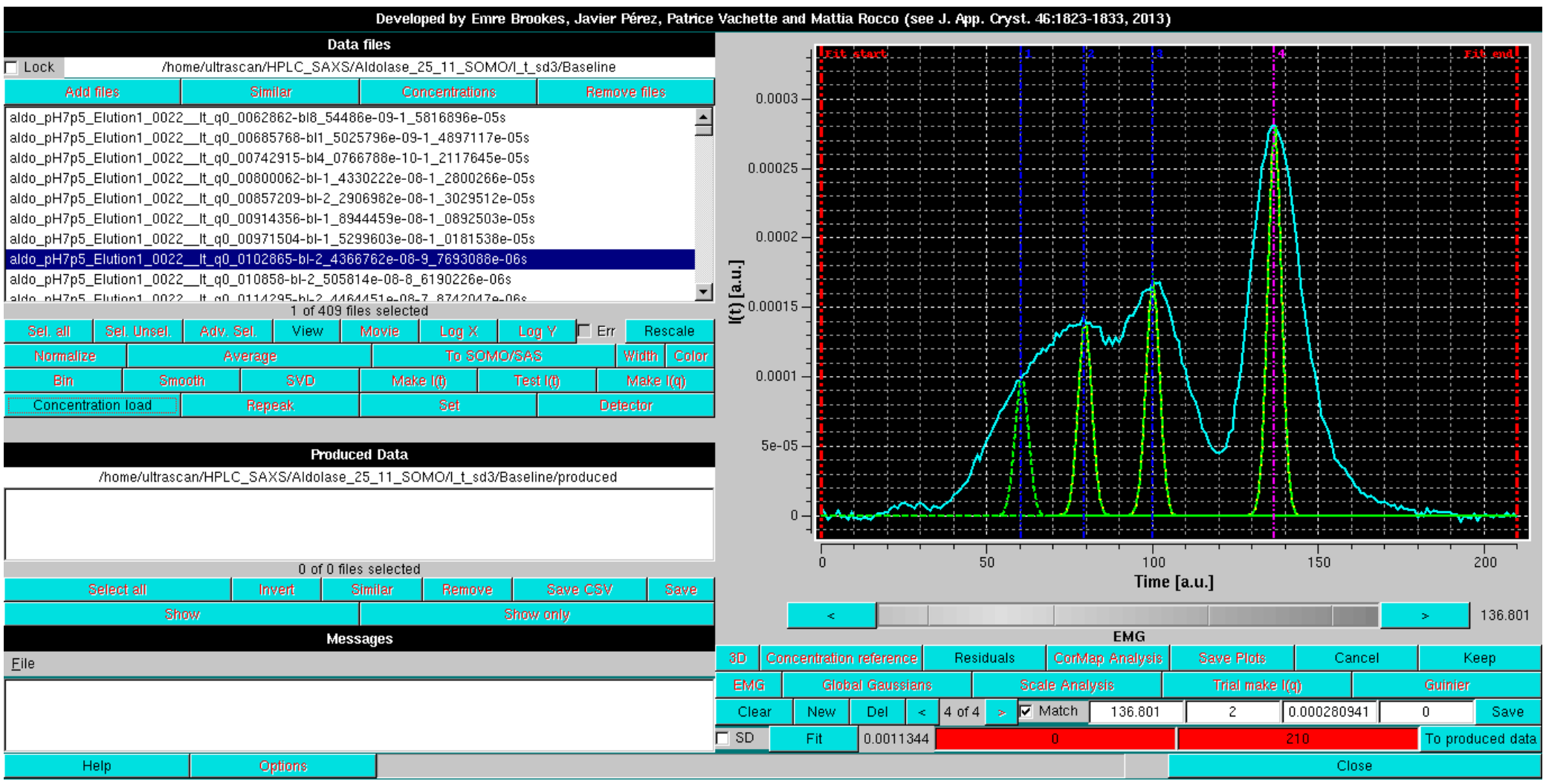
The type of distorted Gaussian to be utilized is chosen in the Options panel accessed by pressing the ***Options*** button:



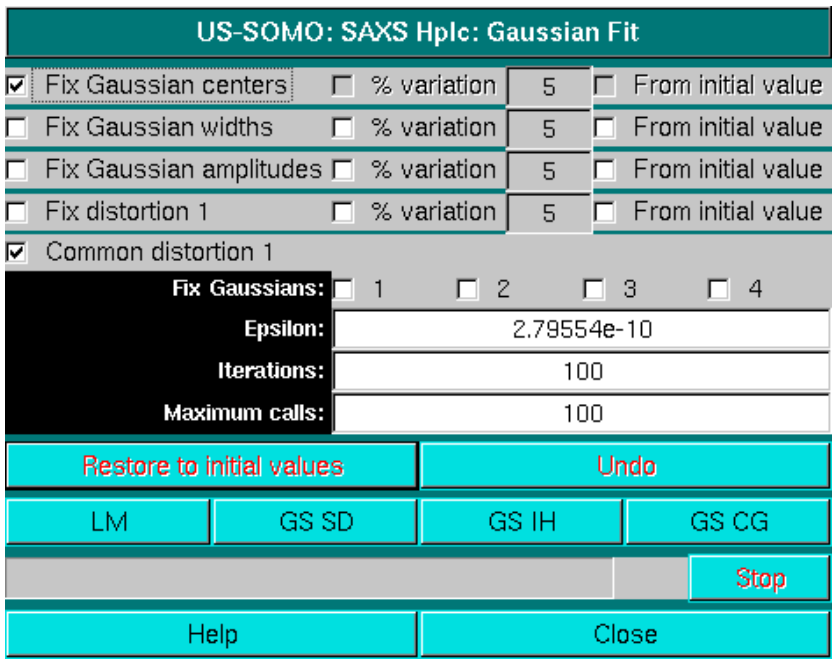
We will start with the Exponentially modified Gaussian (EMG) function. After selecting it and returning to the ***HPLC-SAXS*** main panel, a ***EMG*** button will replace the default ***Gaussians*** button. Pressing it will bring up the EMG settings:



Since an ***SVD*** analysis (see here) on this dataset (not shown) indicated that four components were at least needed to describe it, four EMG Gaussians are initally positioned:



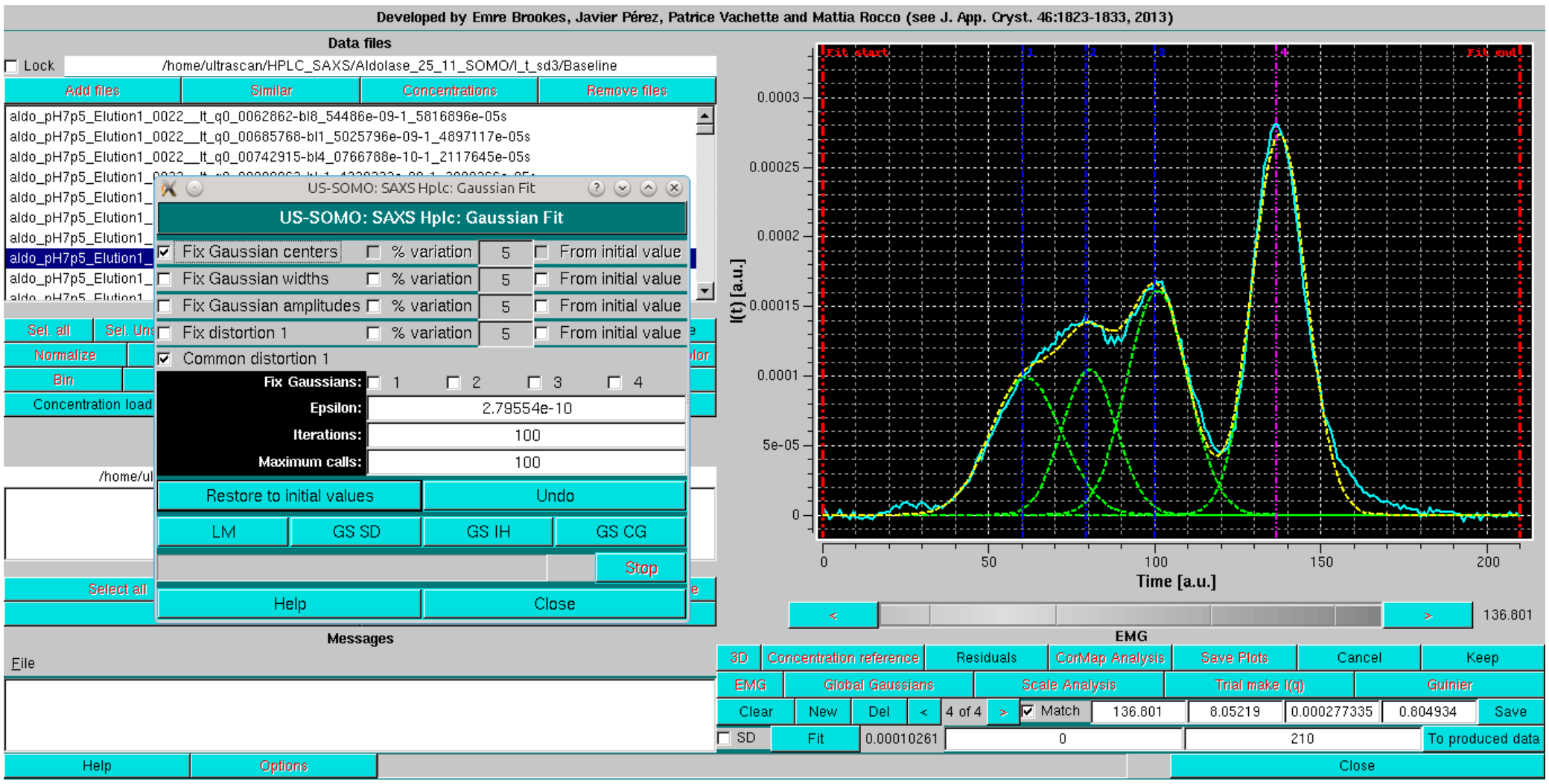
Note that four fields are now present in the third commands row. The first three are the center, width and amplitude of each Gaussian, as in normal Gaussian operation, while the fourth is the EMG Gaussian distorsion (set to 0 at the beginning). All other commands are identical as for normal Gaussian operations. Once the intial set of EMG Gaussians is positioned, pressing ***Fit*** will bring up again the ***Gaussian Fit*** module:



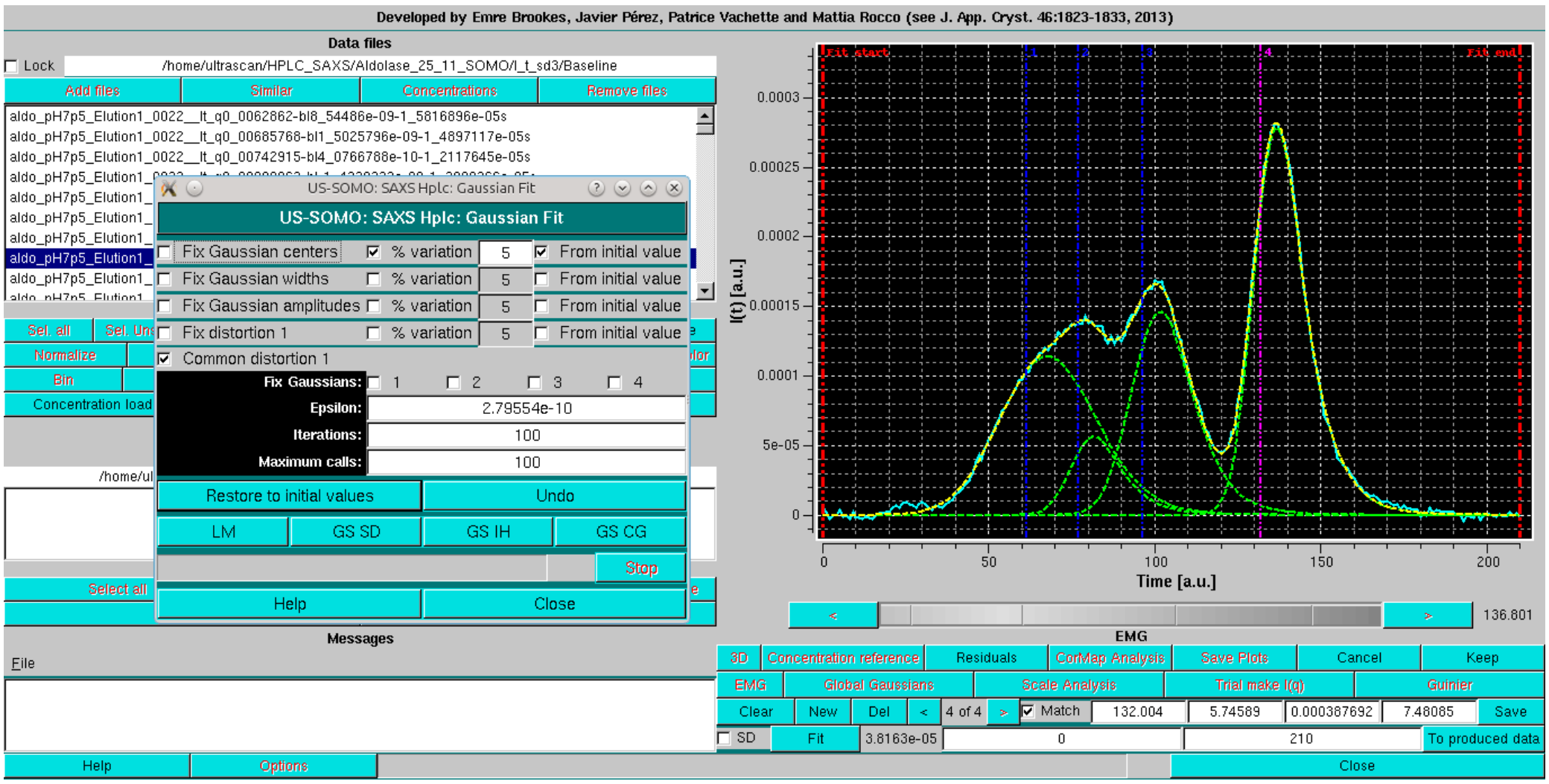
Note that in respect to the normal Gaussians operation, there is an additional distorsion field with its checkboxes (*Fix distortion 1*, *% variation*, and *From initial value*), and a *Common distorsion 1*

checkbox. The latter is selected by default, because it is assumed that similar species will have similar distorsions on eluting from the column. This makes the Gaussian fitting more robust. If necessary, once an initial round of fitting is performed, this constraint can be released, to verify if any further improvements are possible while still keeping resonable peak shapes for all Gaussians.

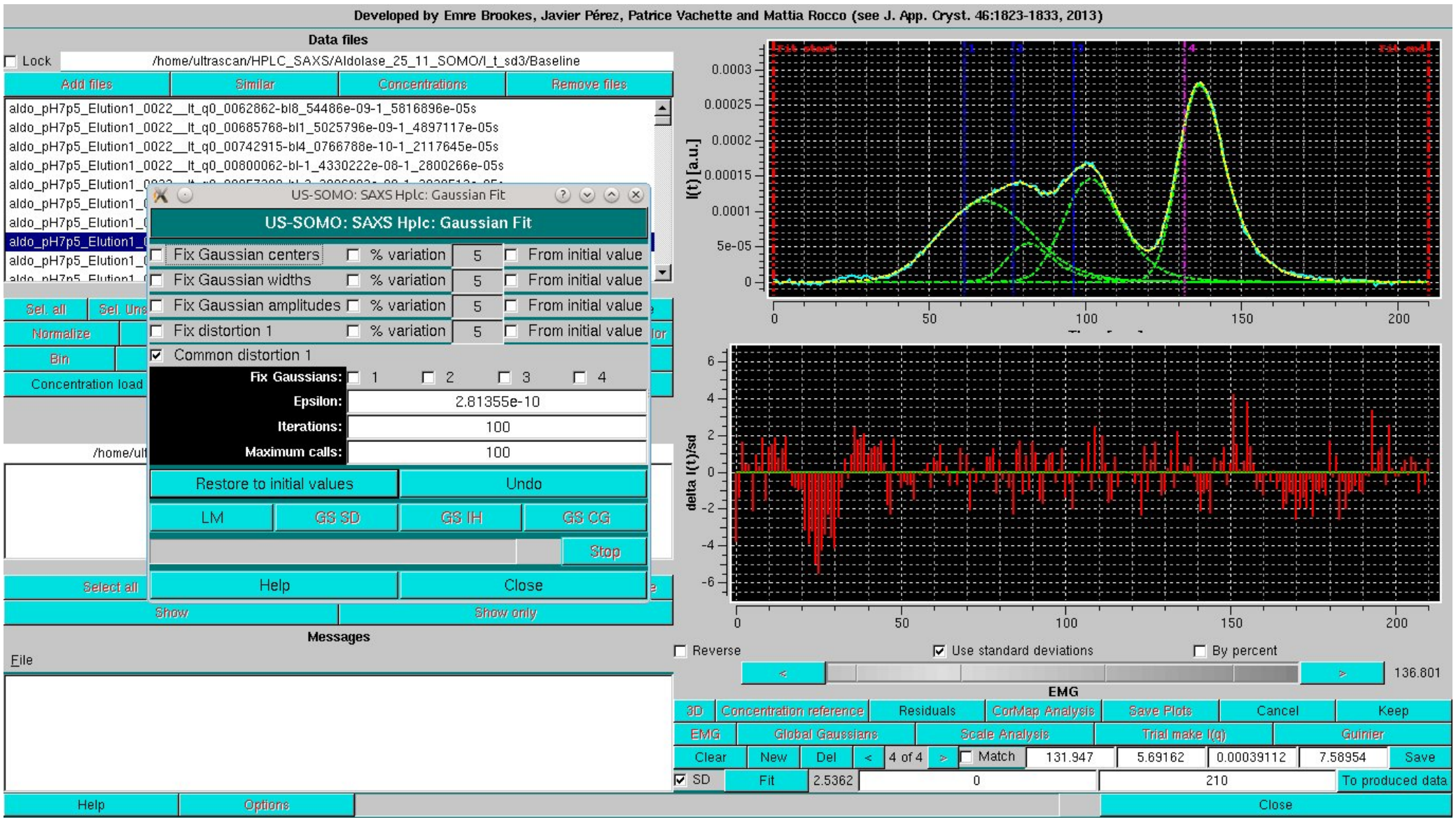
As with normal Gaussain operation, it is advisable to do a first fit while keeping the *Fix Gaussian centers* checkbox selected:



Followed by a round with the centers restrained by the *% variation 5 from initial value*:

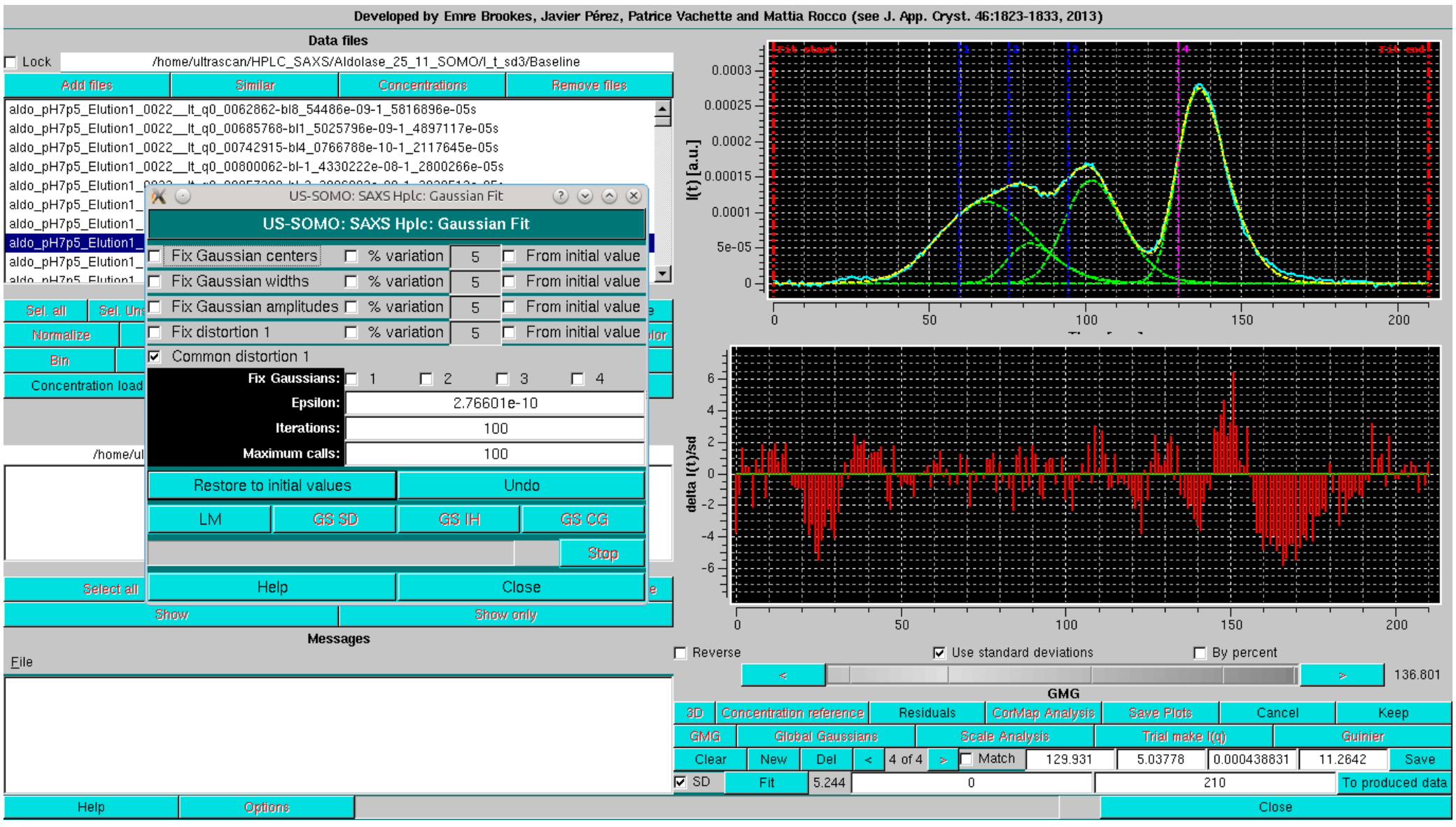


Bringing in the SD and releasing all constraints but the *Common distorsion* produces a slightly improved fit:

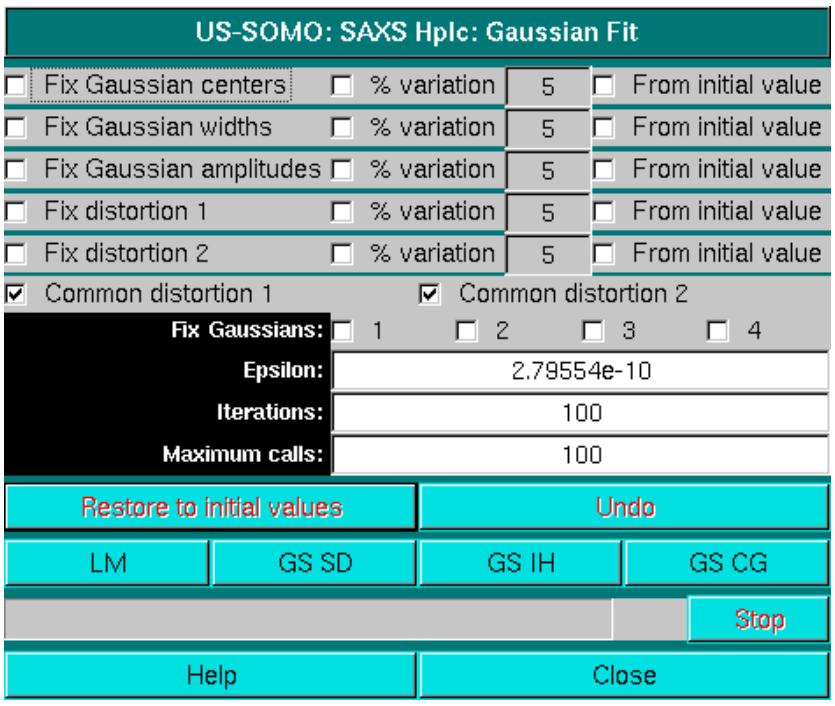


However, the cost paid to achieve this apparently quite satisfactory fit is that the contribution of peak #1 appears to be exaggerated.

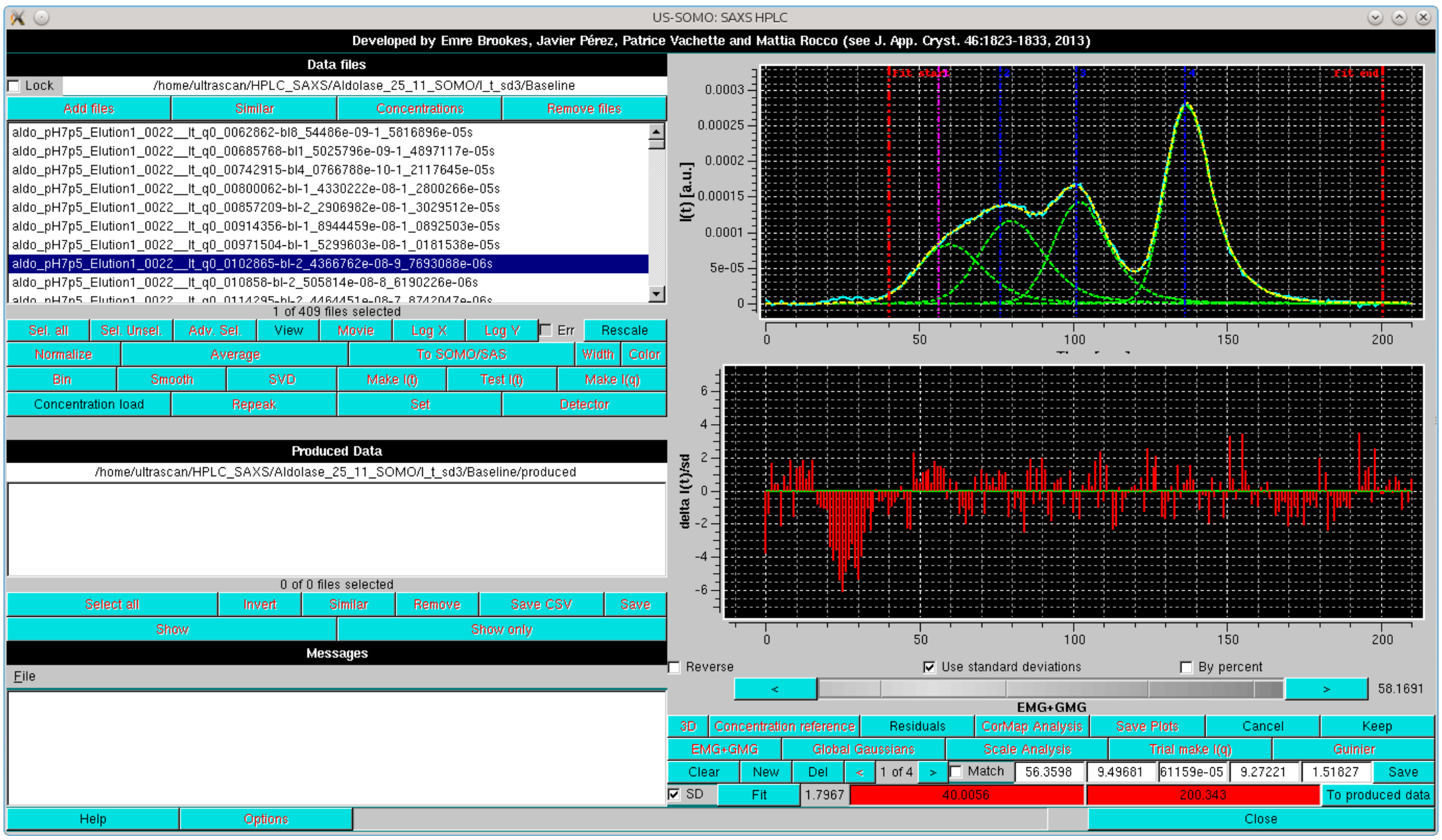
We can perform the same analysis using the ***Half-Gaussian modified Gaussian*** function:



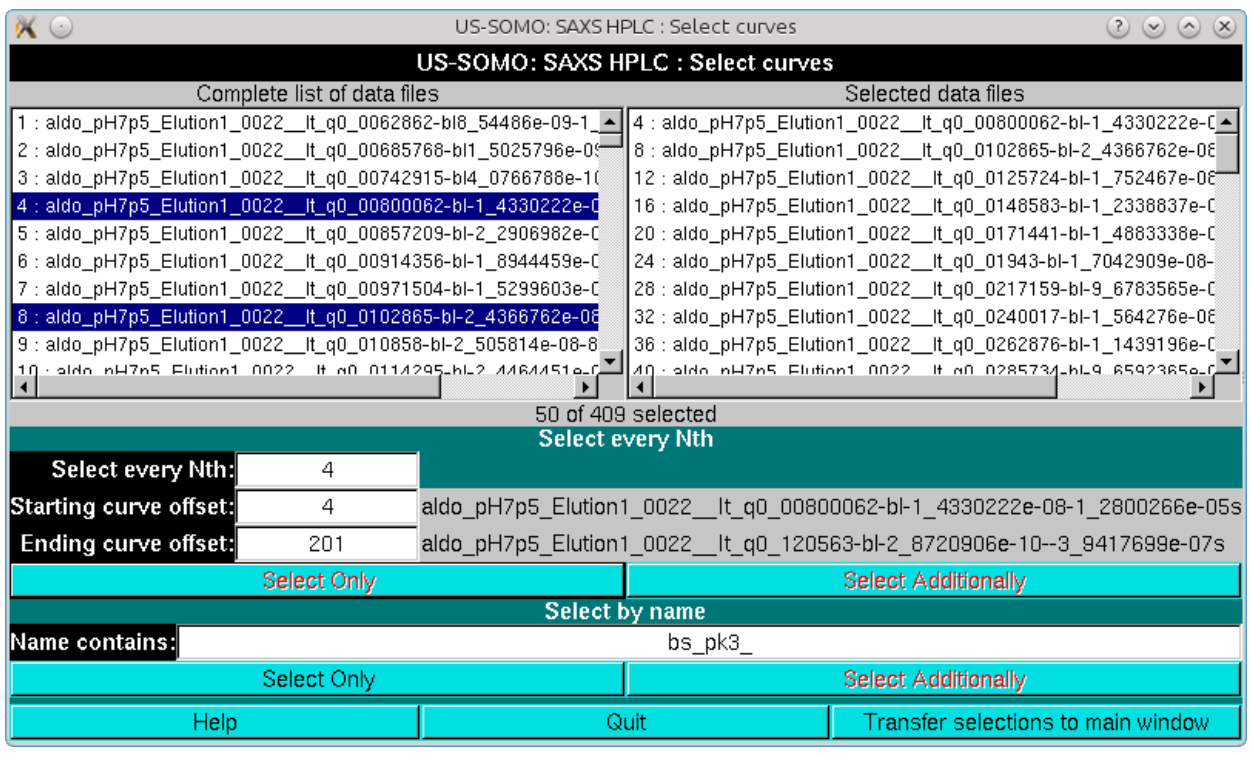
As can be seen from the fit χ^2 (next to the ***Fit*** button), this function performs worse for this dataset. A function combining the EMG and GMG Gaussians (***EMG+GMG***) can be also tested. The corresponding ***Fit*** module will present extra fields:



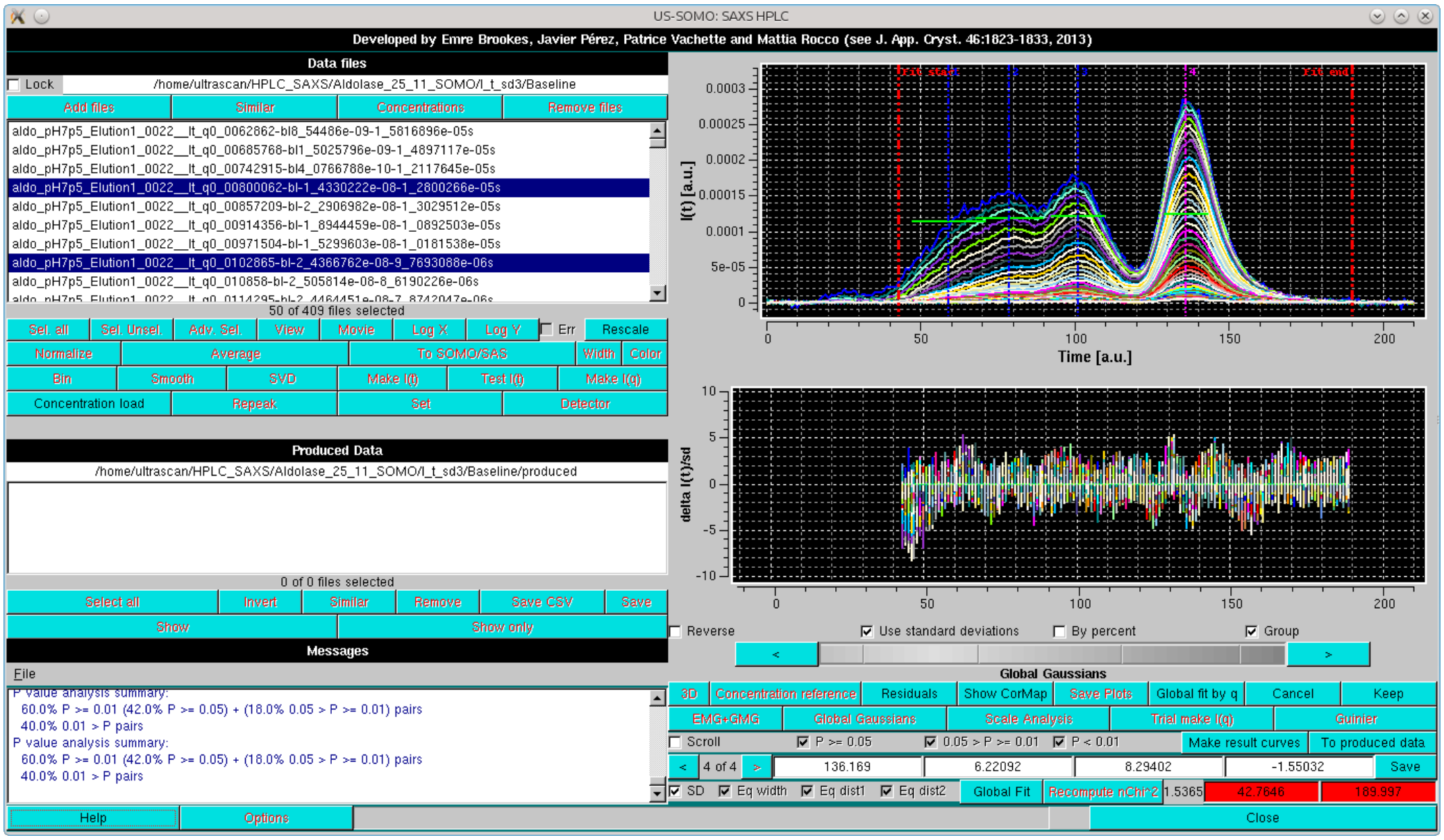
see here for a complete description description of the ***Fit*** module. In addition, we will also restrict the fitting region, to help improve the fitting. The results of this EMG+GMG fitting are shown below:



As can be seen, there is a substantial improvement, mainly due to the exclusion of the small bump before the first peak. We will then proceed with this set, first by doing ***Global Gaussians*** on a subset:



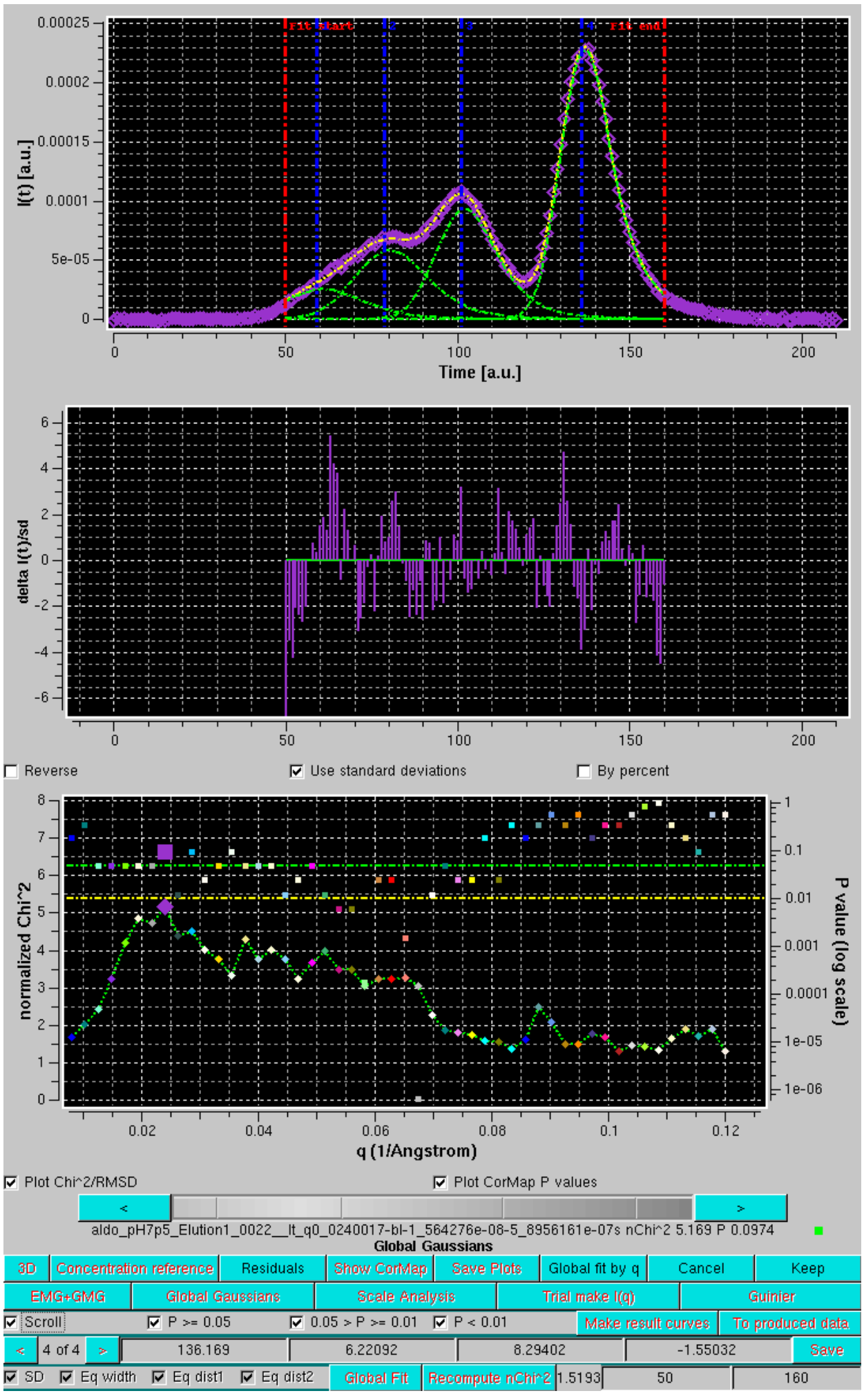
and then proceeding with ***Global fit***:



Here, already by looking at the ***Residuals***, the fit appears to be improved, and the apparently larger residuals especially around the main peak are due to the very low SD associated with the data, amplifying the discrepancies. The goodness of the fit can be further checked by first resticting again the fit limits, then recomputing the residuals by pressing ***Recompute nChi^2***, and subsequentely pressing ***Global fit by q***:

This will bring up an additional plot where the normalized χ^2 (diamonds connected by a line) and the pairwise CorMap *P*-values (squares) are plotted as a function of the *q*-value. In the *Global fit by q* graph it is possible to visualize either one of or both the two plots, by selecting/deselecting their respective checkboxes positioned just below it (*Plot Chi^2/RMSD* and *Plot CorMap P values*). Note that in the image above, where both plots are shown, their respective *y*-axis scales have been manually modified to allow a better visualization of each plot. The dashed green and yellow horizontal lines mark the usual cut-off *P*-values ($P \geq 0.05$, above the green line; $0.05 > P > 0.01$ between the green and yellow lines; $P < 0.01$, below the yellow line).

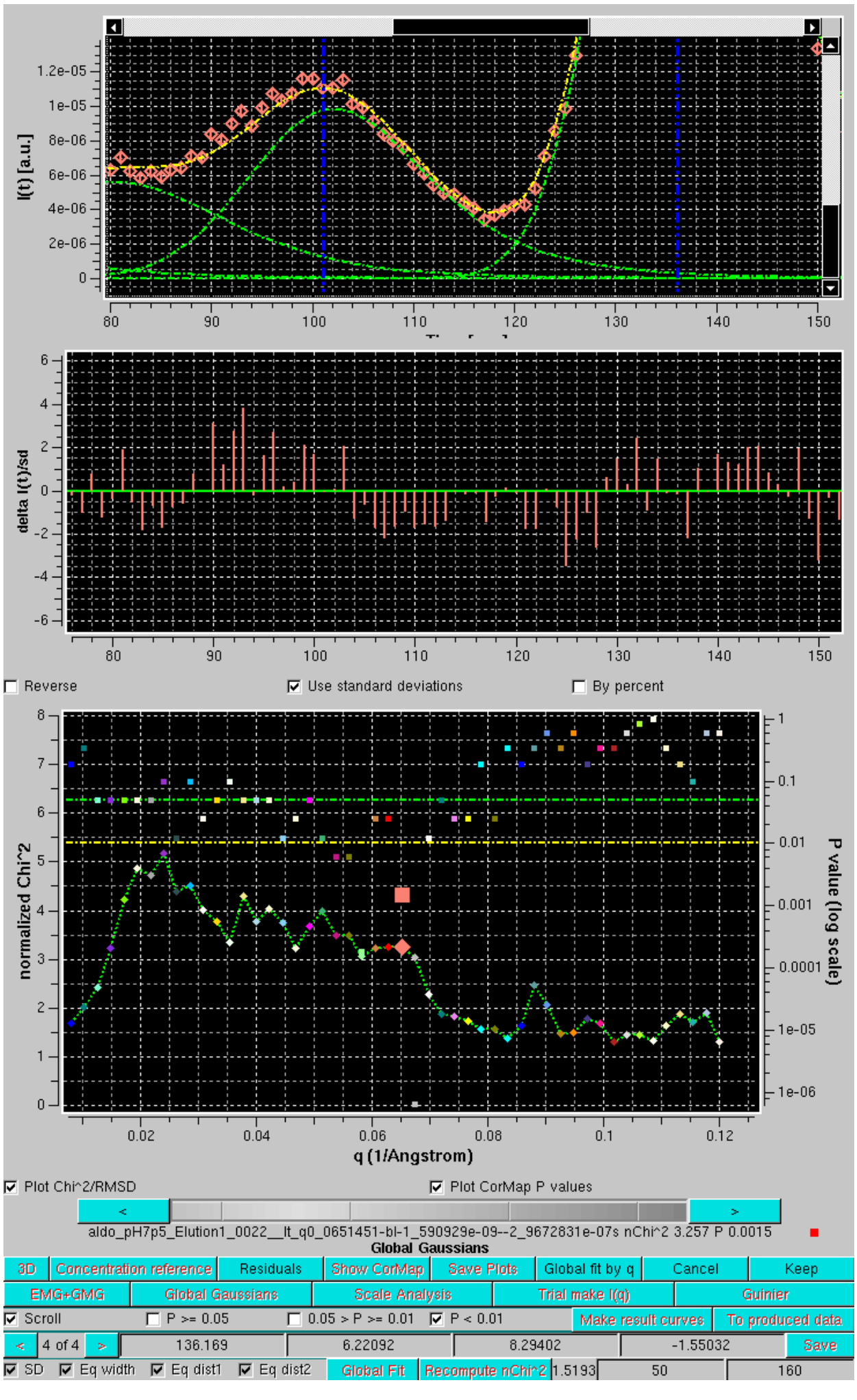
The correlation between the goodness-of-fit indicators and the distribution of the residuals can be examined for each original/fit *I*(*t* vs. *t* pair by selecting the *Scroll* checkbox:



The current chromatograms pair is highlighted in both plots by an enlarged symbol (purple square in this case). Scrolling is performed by either using the grey-scale bar-wheel, or by clicking on the the "<" and ">" buttons placed at its sides. By selecting/deselecting the three checkboxes next to the *Scroll* checkbox (*P >= 0.05*, *0.05 > P >= 0.01*, *P < 0.01*), only the subset(s) whose *P*-values are within those of the selected chechbox(es) will by scrolled.

Note how here to a relatively "bad" χ^2 value (5.169) corresponds a "good" *CorMap* value (*P* = 0.0974). This indicates an essentially random residual distribution, while the high χ^2 value is mainly due to the very low SD associated with the data, since the fit appears also to be quite good (top graph).

Conversely, if we examine a *q*-value where the χ^2 is better (3.257) but the *P*-value is bad (0.0015), we can see by zooming on the inflexition between the 3rd and 4th peaks that the latter is mainly due to a stretch of *I*(*t*) experimental values (salmon diamonds) sligthly below the EMG+GMG fit curve:



After accepting the Global Fit, the EMG+GMG Gaussians are then propagated to all chromatogram by first selecting all chromatograms amd then pressing ***Global Gaussians***:

where the goodness of the reconstruction can be appreciated in both the *Reduced Residuals* and the *Global fit by q* plots.

Save and ***Keep*** can then be sequentially pressed to store and accept the global Gaussian results. After non-symmetrical Gaussian decomposition, the ***Trial make I(q)*** procedure can be launched to further test the results, as described in the main *Help* pages.

---

www contact: Emre Brookes

SOMO HPLC-SAXS Module SVD Utility:

Last updated: April 2016

This utlity was developed to perform single-value decomposition (SVD; e.g., Williamson et al., Biophys J. 94, 4906-4923, 2008) on a set of *I(q)* vs. *q* data, which can come from a HPLC-SAXS experiment, or from any other SAS type of data, like a concentration series.

A single SAS experimental dataset is typically represented as *I(q)*, where *q* is a grid of points. A sequence of *n* $I(q)_{t(j)}$ on the same *q*-grid can be assembled into a $m \times n$ matrix $\boldsymbol{I} = [\boldsymbol{I}_{ij}] = [I(q)_{t(1)}, I(q)_{t(2)}, ..., I(q)_{t(n)}]$. Each column of ***I*** contains an *I(q)* curve for a specific *t* and each row contains an *I(t)* curve for a specific *q*. If standard deviations of the experimental data are available, these can be analogously placed in a matrix ***S***. If a synchronized concentration dataset *C(t)* is available, it can be added to ***I*** as an additional row.

In a SVD analysis, if ***I*** is the matrix containing the original data, then:

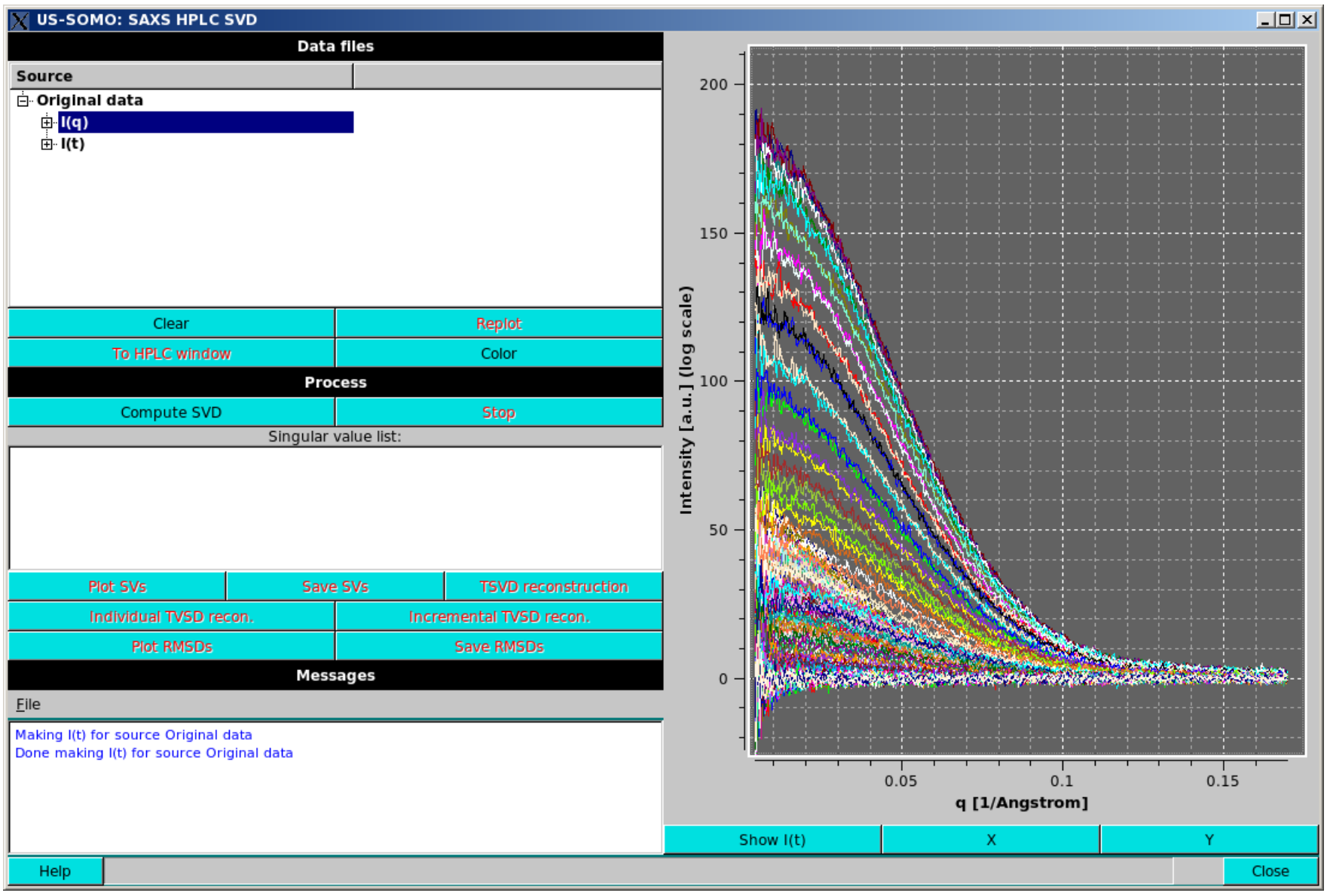
$$\boldsymbol{I} = \boldsymbol{USV^T}$$

where ***U*** is an orthogonal $m \times m$ matrix, ***S*** is a diagonal $m \times n$ matrix, and $\boldsymbol{V^T}$ is an orthogonal $n \times n$ matrix. The elements of the diagonal of ***S*** are the singular values. Reconstructing an approximation of the matrix ***I*** proceeds by setting some diagonal elements of ***S*** to zero, forming ***S'*** and performing the multiplication

$$\boldsymbol{I'} = \boldsymbol{US'V^T}$$

SVD can be performed either on the original or on the baseline-subtracted I(q) vs. q data or subset of data (if significant baseline drift occurs, the SVD will try to fit also that part of the signal).

After selecting the data, pressing the ***SVD*** button in the **HPLC-SAXS** module will open a new window:



The top left box labelled **Data files** will contain a list of the data set in an expandable format. The first set will be labelled "Original data". Opening the item will show its contents, "I(q)" and "I(t)", which can be further expanded to show/select the individual curves in the data set.

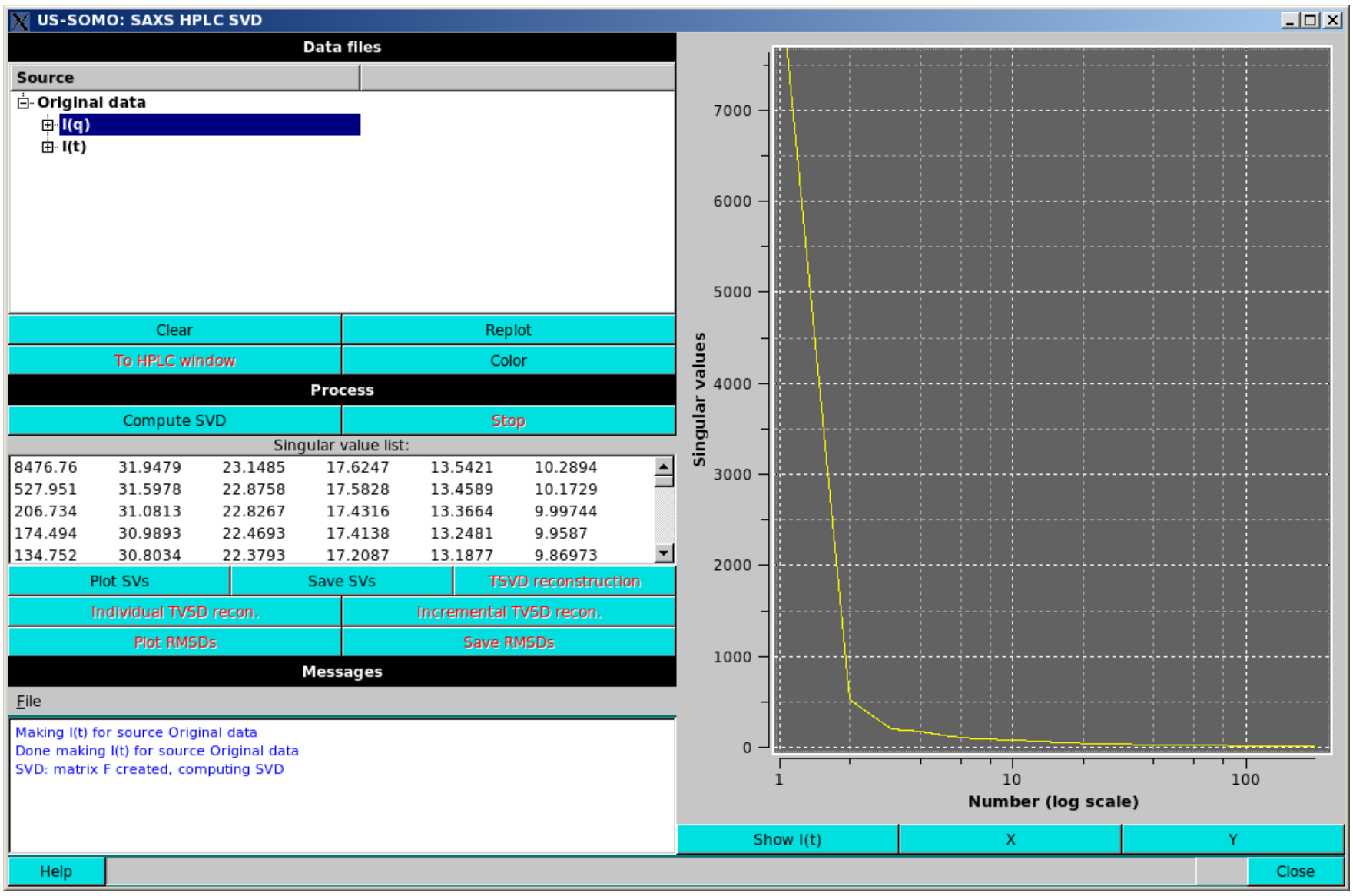
Replot will become active when a selection is changed from what is currently plotted; pressing it will refresh the plot display. For example, selecting "Original data" at the top level and pressing ***Replot*** will plot the entire dataset, but only in the "I(q)" or "I(t)" mode.

TO HPLC window will transfer the selected dataset back to the **HPLC-SAXS** module main window.

Show I(t) or ***Show I(q)*** toggle button below the plot window will show the *I(t)* vs. *t* or *I(q)* vs. *q* view of the data and automatically replot.

Color will rotate the plot colors based upon a pre-defined palette.

When a single data set or sub-selection of individual *I(q)* curves from a single dataset are selected and the plot window is in "I(q)" mode, the ***Compute SVD*** button in the **Process** box becomes active.



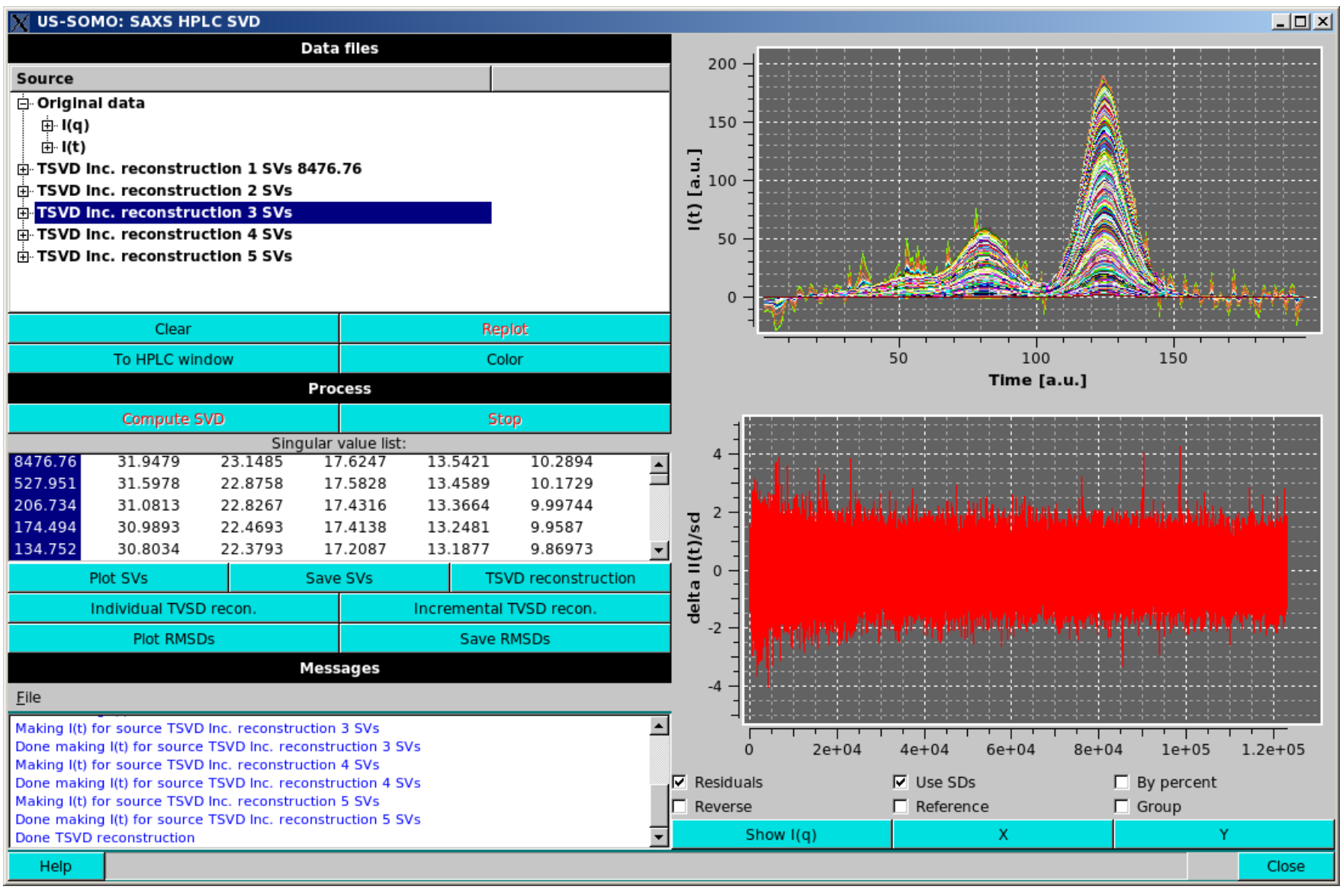
Pressing ***Compute SVD*** will compute the singular value decomposition (Lawson & Hanson, Solving least squares problems. SIAM, Philadelphia, 1995). The singular values (SVs) will be sorted in descending order and placed on the screen in the *Singular value list* window.

Plot SVs will plot the SVs in the plot area, by default in a linear vs. log scale (see above).

The axes scales can be toggled between logarithmic and linear by pressing the ***X*** or ***Y*** buttons below the plot window.

Save SVs will save the SVs to a file. Giving the filename a ".csv" extension will result in a comma separated output, otherwise, the output will be "TAB" separated.

Selecting any set of SVs in the *Singular value list* by clicking on it will activate three more buttons: ***TSVD reconstruction***, ***Individual TSVD recon.***, and ***Incremental TSVD recon.***.



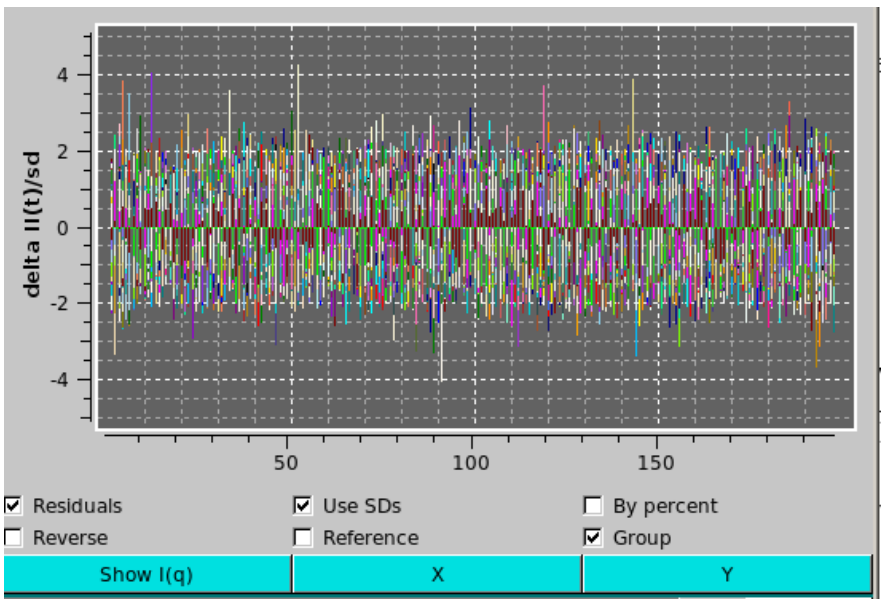
TSVD reconstruction will generate a new dataset in the **Data files** section consisting of the reconstruction of the data based upon the selected SVs. "TSVD" means a truncated SVD reconstruction (Aster et al., Parameter Estimation and Inverse Problems. Elsevier Academic Press, 2005), which formally should be computed on the numerically highest SVs, but here we are using the term loosely, to mean reconstruction on any subset of the SVs. The resulting dataset can be selected, expanded and/or plotted identically to the original data. Expanding the TSVD data will show "I(q)", "I(t)", "SVs used" expandable subsections and also the root mean squared deviation over the number of points (RMSD) of the expansion and the name of the reference dataset for the reconstruction.

Individual TSVD recon. will take the selected SVs and produce a TSVD reconstruction for each value selected individually, resulting in multiple TSVD reconstructions in the **Data files** section.

Incremental TSVD recon. will take the selected singular values and produce a TSVD reconstruction for the first value selected, then for the first and second values selected, etc., until all the selected values are included in a reconstruction, again resulting in multiple reconstructions in the **Data files** section, as in the example shown above. Note that the display has been swithced to the "I(t)" mode, where the goodness of the reconstruction can be better appreciated than in the "I(q)" mode.

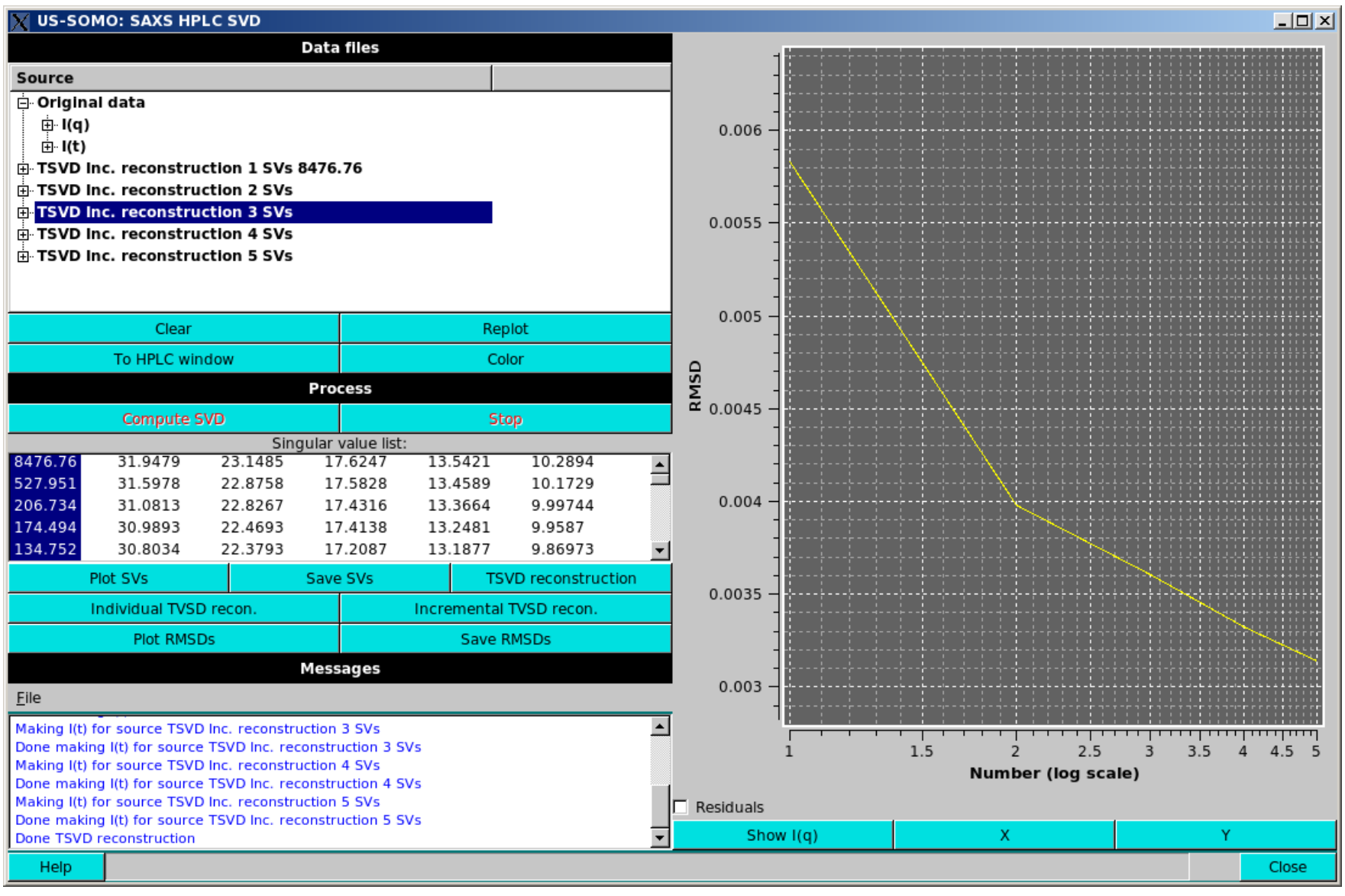
When ***Individual TSVD recon.*** or ***Incremental TSVD recon.*** are selected, the *residuals* checkbox under the plot appears. Selecting this checkbox will plot the residuals of the reconstruction vs. the reference dataset in a new plot area below the main plot area. When residuals are displayed, additional checkboxes will be available under the plot. These include:

- *Use SDs*, which turns on division of the residuals by the SD of the reference dataset and will not be available for data without SDs (as in the example shown above). Note that in the default mode, the residuals are displayed in a linear contiguous manner for each original dataset;
- *By percent*, which displays the differences by percent;
- *Reverse* which reverses the Y-axis around zero;
- *Reference*, which toggles on/off the display of the reference data in the main plot window;
- *Group*, which toggles grouping of the residuals between the linearly contiguous manner and superimposed (as shown below).



Once an individual or incremental reconstruction is computed, two more buttons activate:

Plot RMSDs and ***Save RMSDs***, allowing plotting and saving analogous to the ***Plot SVs*** and ***Save SVs*** mentioned previously.



As mentioned above, any reconstructed dataset can then be added to the **US-SOMO/HPLC-SAXS** module by selecting and pressing the ***To HPLC window*** button. Note only the "I(q)" or "I(t)" data will be added depending on the plot mode.

As an example, suppose one wants to determine the number of components present in a set of *I(q)* vs. *q* curves. After bringing them into the **SVD** module as described above, the SVD can be then computed. By looking at the SVs plot, one can evaluate that at most *N* singular values seem reasonable to reconstruct the dataset. One would then select the numerically largest *N* values in the SV list and run an incremental reconstruction. Subsequently, each reconstructed dataset could be compared by RMSD and visually to determine the effect of adding additional singular values to the reconstruction to assist the determination of the minimum number of singular values required to accurately reconstruct the original data. Another check would be to run the individual reconstruction on the same set of selected singular values and inspecting the individual datasets visually (preferable via I(t) plots) to see if there seems to be signal present in reconstructions past a minimum number of singular values. In this way, the **US-SOMO/HPLC/SVD** module can be used to approximate the number of independent components present in a HPLC experimental dataset or even a concentration series.

---

www contact: Emre Brookes

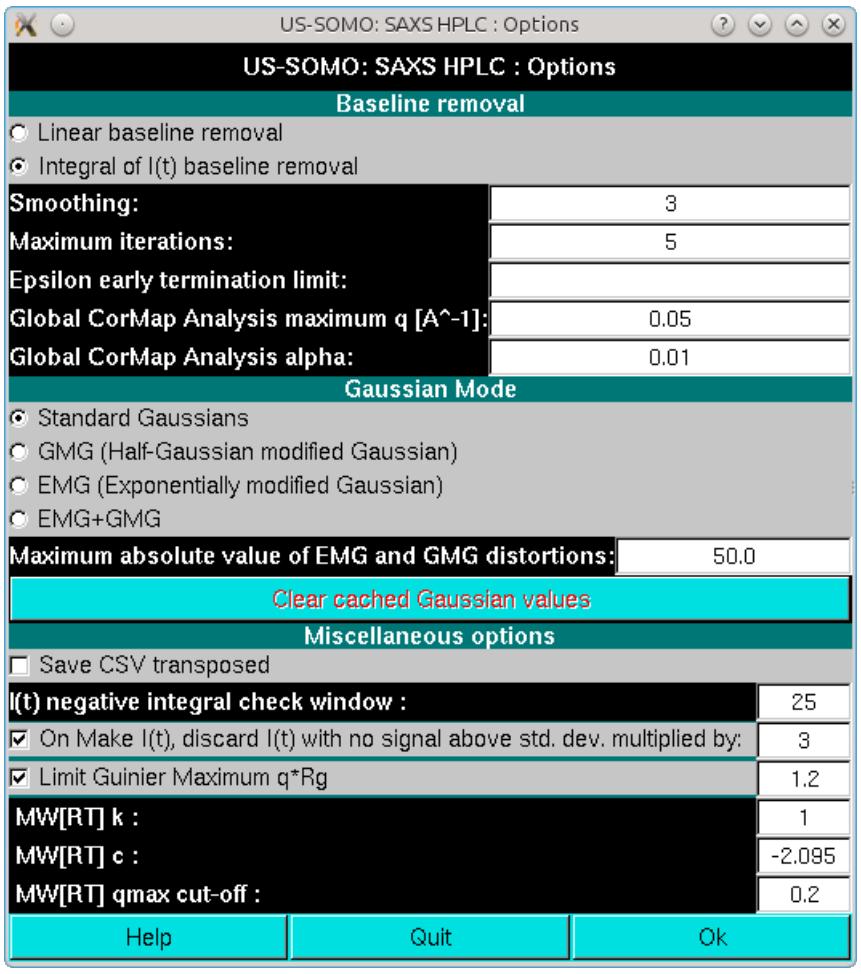
The latest version of this document can always be found at:

http://somo.uthscsa.edu

Last modified on April 18, 2016.

SOMO HPLC-SAXS Module Options Panel

Last updated: May 2016



The first set of options deals with ***baseline removal*** tools. By ***default***, the **US-SOMO *HPLC-SAXS*** module will utilize an ***Integral of I(t) baseline removal*** tool. The method is based upon the assumption that capillary fouling deposits are formed in proportion to the sample intensity while exposed to the beam (and that the buffer is not responsible). We have developed a mathematical model of this condition and implemented it as the integral baseline procedure.
This is an iterative procedure and the numberof steps can be set in the *Maximum iterations:* field (***default: 5*** iterations). The program will generate a baseline at each iterative step, which will be showed when testing the results.
As part of this procedure, the *epsilon* value corresponds to the intensity fouling deposits constant and an early termination criteria can be defined here which will stop the iterative procedure when the difference in *epsilon* between iterations is less than or equal to the *epsilon early termination limit* field (***default: none***).
The *Smoothing* field controls the size of a Gaussian smoothing kernel of 2n+1 points (***default: n=3***) that will be applied to each $I(t)$ chromatogram prior to Integral Baseline computation, to avoid large oscillations especially at low q values which might cause problems. The Integral Baseline is nevertheless applied to the **original** $I(t)$ chromatograms, not to the smoothed data!
The *Global CorMap Analysis maximum q [A^-1]* field defines the upper q value limit (***default: 0.05 Å^{-1}***) for the CorMap analysis of datasets (see here).

For special needs or very minor baseline correction requirements, the original linear baseline implementation (Brookes et al., J. App. Cryst. 46:1823-1833, 2013) is available by selecting the ***Linear baseline removal*** checkbox. A description of the linear baseline removal tool can be found here)

The next set of options deals with the ***Gaussian mode***. Four alternative checkboxes are available:

- ***Standard Gaussian***. This will generate symmetrical Gaussians (***default option***).
- ***GMG (Half-Gaussian modified Gaussian)***. This will generate skewed Gaussians according to this equation:.

$$y = \frac{a_0 \exp\left(-\frac{1}{2}\frac{(x-a_1)^2}{a_3^2+a_2^2}\right)\left[1+\operatorname{erf}\left(\frac{a_3(x-a_1)}{\sqrt{2}a_2\sqrt{a_3^2+a_2^2}}\right)\right]}{\sqrt{2\pi}\sqrt{a_3^2+a_2^2}}$$

where a_0, a_1, a_2, and a_3 are the *area*, *center*, *width*, and *distorsion*, respectively, of the half-Gaussian modified Gaussian(s).

- ***EMC (Exponentially modified Gaussian)***. This will generate skewed Gaussians according to this equation:.

$$y = \frac{a_0}{2a_3}\exp\left(\frac{a_2^2}{2a_3^2}+\frac{a_1-x}{a_3}\right)\left[\operatorname{erf}\left(\frac{x-a_1}{\sqrt{2}a_2}-\frac{a_2}{\sqrt{2}a_3}\right)+\frac{a_3}{|a_3|}\right]$$

where a_0, a_1, a_2, and a_3 are the *area*, *center*, *width*, and *distorsion*, respectively, of the exponentially modified Gaussian(s).

- ***EMG+GMG***. This will generate skewed Gaussians according to this equation:.

$$y = \frac{a_0}{4a_3}\exp\left(\frac{2a_1a_3-2a_3x+a_2^2}{a_3^2}\right)\operatorname{erfc}\left(\frac{a_1a_3-a_3x+a_2^2}{\sqrt{2}a_2a_3}\right)+\frac{a_0}{2\sqrt{2\pi}\sqrt{a_2^2+a_4^2}}\exp\left(-\frac{1}{2}\frac{(a_1-x)^2}{a_2^2+a_4^2}\right)\operatorname{erfc}\left(\frac{a_4(a_1-x)}{\sqrt{2}a_2\sqrt{a_2^2+a_4^2}}\right)$$

where a_0, a_1, a_2, a_3, and a_4 are the *area*, *center*, *width*, *distorsion 1*, and *distorsion 2*, respectively, of the exponentially + half-Gaussian modified Gaussian(s).

The ***Maximum absolute value of EMG and GMG distortions:*** field sets limits to the distortions to avoid unreasonably skewed Gaussians (***default: 50***).

The ***Clear cached Gaussian values*** button allows to clear all Gaussian data produced during the current session.

In the ***Miscellaneous options*** section there is the option of saving the *.csv files in transposed format, by selecting the ***Save CSV transposed*** checkbox.

It is followed by the ***I(t) negative integral check window*** field, where the sliding window size for the negative integral test can be set (***default: 25***). As described previously, this is the size of a sliding window over adjacent frames. If the sum of $I(t)$ values within the window is less than the negative of sum of the corresponding SD values, the $I(t)$ curve will be identified and a warning will be issued. Such regions can be indicative of not optimal buffer subtraction, and might cause problems with the integral baseline correction.

The next line contains the checkbox for the ***On Make I(t), discard I(t) with no signal above std. dev. multiplied by:*** and its associated field (***default: checked*** with ***3*** as the multiplier value). This tests whether or not there is any point in the $I(t)$ curve where the $I(t)$ value is greater than the SD of the point of $I(t)$ multiplied by the value in this field. If the test fails, the $I(t)$ curve is assumed to contain no signal and is dropped. A ***Warning*** message appears in an appropriate pop-up messagebox listing the first 20 occurences and how many more were found.

The ***Limit Guinier Maximum q*Rg*** allows to set this limit to a specified value (***default: 1.1***) when performing the linear regression in the Guinier analysis to calculate the R_g and $I(0)$ from the $I(q)$ vs q datasets in the ***Test I(q)*** module (see the corresponding section in the main Help).

Finally, the last three fields allow modifying the constants for the approximate calculation of the molecular weight as described by Rambo and Tainer (Accurate assesment of mass, models and resolution by small-angle scattering. Nature 496:477-481, 2013). The first two fields contain the constants ***MW[RT]k*** (***default: 1***) and ***MW[RT]c*** (***default: -2.095***), which are the linear fit power law constants as described in eq. (2) of the aforementioned reference and whose default values are taken from Fig. 3, where they were reported for $I(q)$ vs. q data limited to q = 0.3 Å^{-1}. The third field contains the actual q_{max} cut-off value to which the available $I(q)$ vs. q data will be limited (***default: 0.2 Å^{-1}***). The constants default values are for globular proteins, and the cut-off value was found empirically to provide better results. If any of these values is changed, a warning will pop-up:

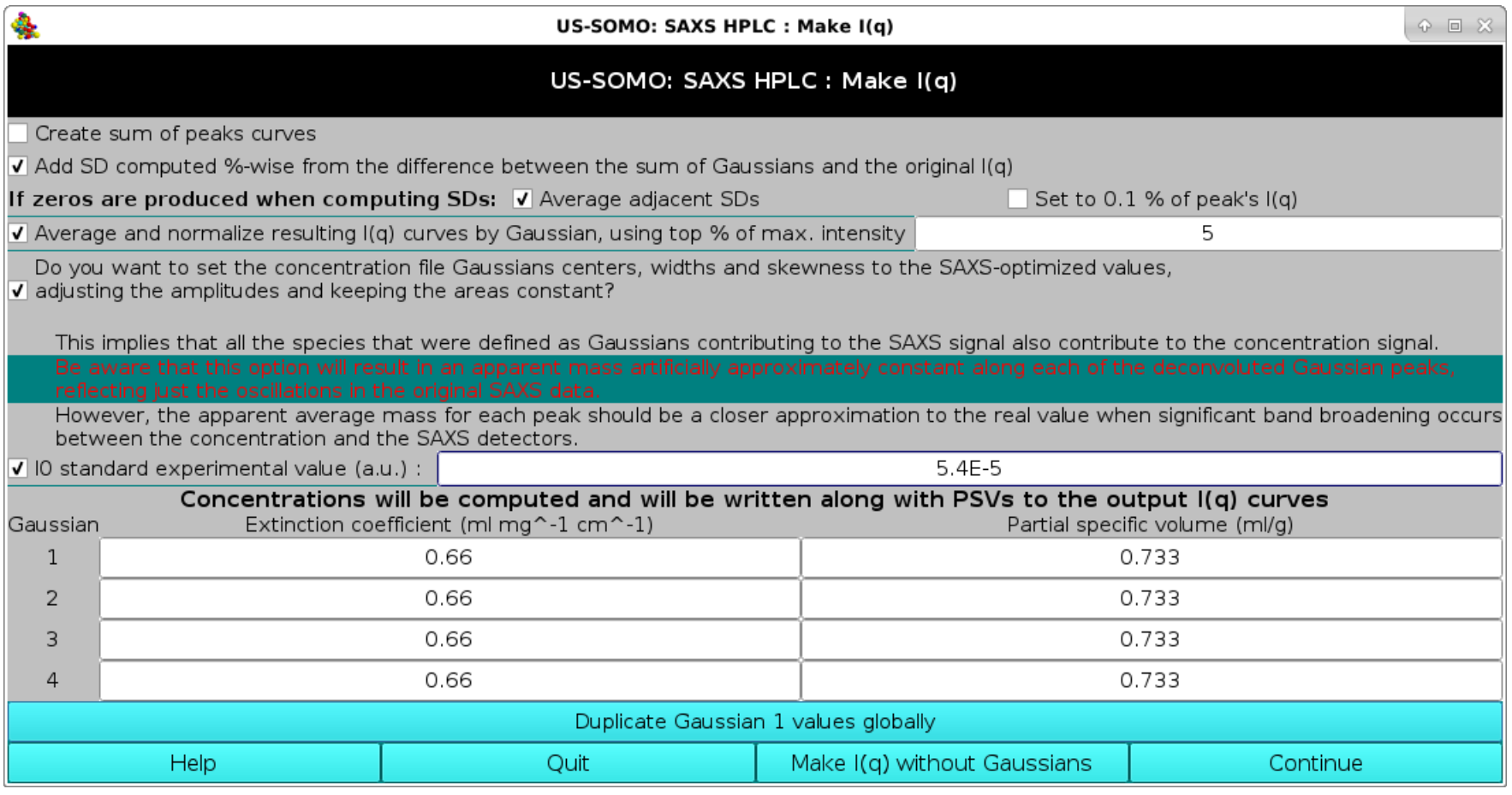
If the values are not reverted to the default values, this pop-up will keep showing up during the current ***US-SOMO*** session every time a molecular weight calculation with the Rambo-Tainer approach is performed, unless the *Do not show this warning again* button is pressed, avoiding the nuisance when using different values for different kind of biomacromolecules (e.g., RNA) or for testing purposes.

---

www contact: Emre Brookes

SOMO HPLC-SAXS Module Make I(q):

Last updated: January 2018



This pop-up panel will appear when the ***Make I(q)*** button is pressed in the main ***HPLC-SAXS*** module.

Three main options are present in this panel:

- The first checkbox *Create sum of peak curves* allows to check that the individual Gaussian will add back up reconstructing the original *I(q)* vs. *q* curves. If checked, their point-wise sum can also be saved, as either *sum(I)* (reconstructed values) and *sum(G)* (pure Gaussians) curves. If baselines were established and subtracted, two sums will be produced, without and with baseline back-addition (default: **unchecked**). If the *integral* baseline procedure was employed, the back-addition will work properly only if the exact integration limits are still present, e.g., if no cropping operation was performed after the baseline correction. For this reason, a **Warning** message is printed below this checkbox.

- *Add SD computed %-wise from the difference between the sum of Gaussians and original I(q)* . In all cases, the original errors associated with each *I(q)* vs. *q* point in each frame will assigned to each point in the resulting decomposed *I(q)* vs. *q* curves for each frame. Alternatively, if this checkbox is checked a new set of additional errors are computed by point-wise calculating the difference between the sum of the Gaussians and the original, baseline-corrected curve. These additional SDs are then assigned %-wise to each point in the decomposed *I(q)* vs. *q* curves, and the new total SDs are calculated at each point by taking the square root of the sum of the square of the original SD plus the square of the Gaussian-derived SD (default: **checked**).

If in computing the SD of the *I(t)* data zeros are produced, an additional line will be present in this panel, as shown above.

The extra line reads *If zeros are produced when computing SDs:* and then offers two options:

- *Average adjacent SDs*. An average between the SD of the previous and following datapoints will be made and assigned to the data point in question (**default option**).

- *Set to 0.1% of peak's I(q)* . The corresponding *I(q)* points will have 0.1% SD.

The third option was recently added (January 2018):

- The checkbox *Average and normalize resulting I(q) curves by Gaussian, using top % of max. intensity"* , allow to automatically select a number of frames whose intensity (as defined by the associated Gaussian) is within a % value of the top frame intensity, as defined in its associated field (default: **unchecked**, with **5%** as the limiting value). All frames selected according to this criterion for each Gaussian-defined peak are first normalized (if an associated concentration file is present; see below), and then averaged. This feature, introduced for the January 2018 release, makes it quite simple to automatically generate averaged datasets after a Gaussian decomposition, without the need to manually check the resulting chromatograms to decide which frames to include in the averages.

If a concentration chromatogram has been associated to the SAXS data, an additional series of options are presented:

- The first deals with the new concentration chromatogram re-shaping routine. The question asked is *Do you want to set the concentration file Gaussian centers, widths and skewness to the SAXS-optimized values, adjusting the amplitudes and keeping the areas constant?*
  A series of considerations and warnings follow. The first states that using this procedure implies that all species that were defined as Gaussians contributing to the SAXS signal also contribute to the concentration signal. If this condition is not met, e.g., when a non-absorbing component is present and a spectrophotometric concentration detector is used, appling this correction can lead to serious mistakes.
  In red then comes a warning: Be aware that this option will result in an apparent mass artificially approximately constant along each of the deconvoluted Gaussian peaks, reflecting just the oscillations in the original SAXS data.
  However, the apparent average mass for each peak should be a closer approximation to the real value when significant band broadening occurs between the concentration and the SAXS detectors.

- *I0 standard experimental value (a.u.):* . A normalization factor from a standard sample to be associated with the data can be also entered here. If this checkbox is checked, its value will be associated in each resulting *I(q)* vs. *q* curve (default: **unchecked**).

- Finally, a calculated concentration can be associated to each of the resulting *I(q)* vs. *q* curves. This is done by entering an extinction coefficient (or a *dn/dc*) for each Gaussian. In addition, a partial specific volume (*psv*) value, needed for the computation of the $<M>_w$, $<M/L>_{w/z}$, and $<M/A>_{w/z}$ by Guinier analysis in the main **US-SOMO *SAS module*** can be also entered here. The module will present as many fields as the Gaussians used to decomposed the data. If all the Gaussians represent a species with the same extinctior coefficient (or *dn/dc*) and *psv*, the values need to be entered only once in the 1st Gaussian fields, and then they can be propagated to all the other fields by pressing the ***Duplicate Gaussian 1 values globally*** button. The concentration value will be used either by the automatic frames-to-be-averaged selector described above, or when the ***Normalize*** button in the main **HPLC-SAXS** module is pressed after dataset selection.

It is also possible to back-generate an *I(q)* vs. *q* set without using the Gaussian decomposition by pressing the ***Make I(q) without Gaussians*** button.

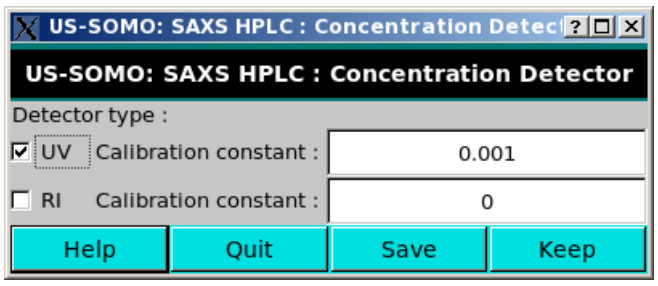
For normal operation, once all fields/checkboxes have been properly set, pressing ***Continue*** will return to the main ***HPLC-SAXS*** window and start the make *I(q)* process.

---

www contact: Emre Brookes

SOMO HPLC-SAXS Module Concentration Detector Selection:

Last updated: October 2013



This utlity will allow to select the type of detector and to enter its calibration constant. Two types of concentration detector are currently supported, UV and refractive index. The used detector can be selected by clicking on either of the two *Detector type:* checkboxes and entering the calibration constant converting their signal in the proper units (absorbance or dn/dc) in the fields to their right:

UV Calibration constant:

RI Calibration constant:

Quit will abort the operation.

Save will permanently associate the type of detector to the ***HPLC-SAXS*** module (it can be changed by re-accessing the ***Concentration detector*** module and chosing again, followed by ***Save***).

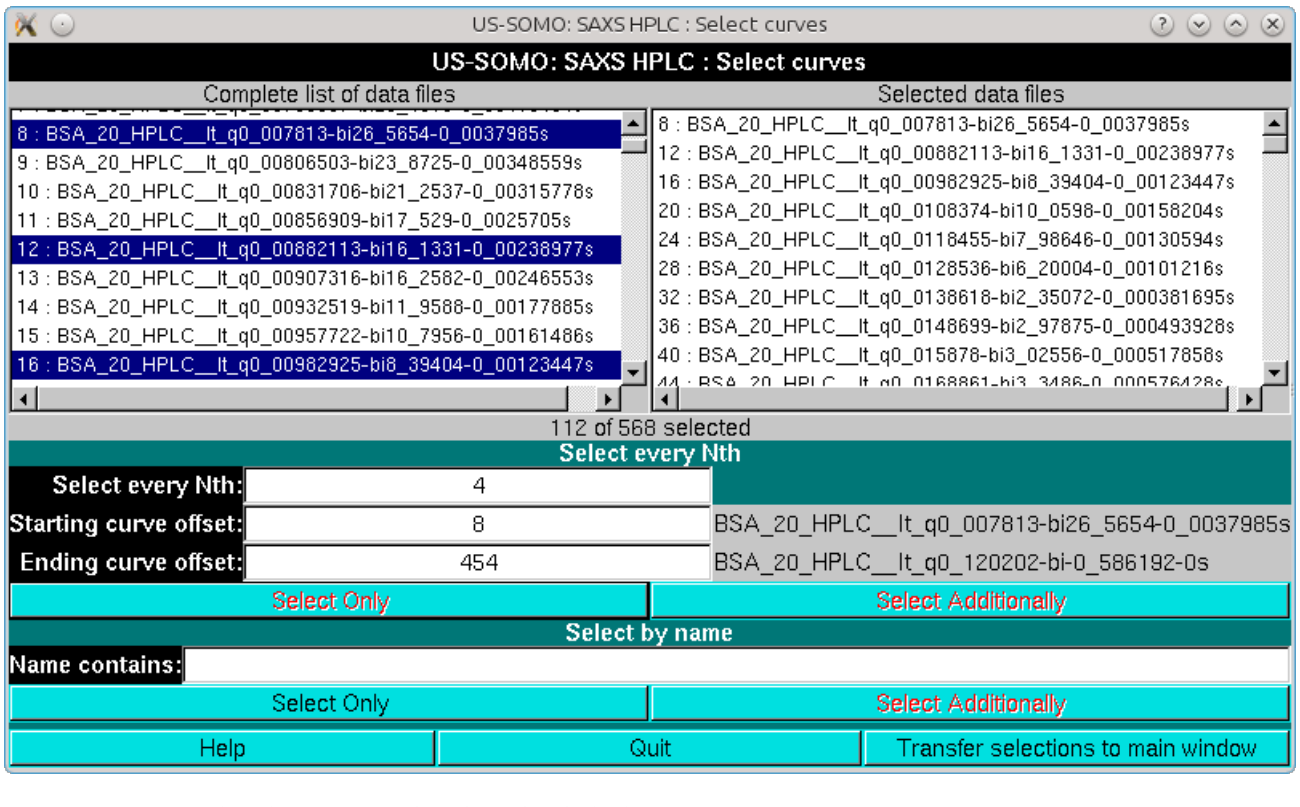
Keep will temporarily associate the type of detector to the ***HPLC-SAXS*** module.

---

www contact: Emre Brookes

SOMO HPLC-SAXS Module Files Selection Utility:

Last updated: April 2016



This utility was implemented to perform efficient sub-selections on already uploaded files.

For instance, in the example shown 1 every 4 files were selected, starting from file 8 and ending at file 454. This was attained using the **Select every Nth** subpanel. First, "4" was entered in the ***Select every Nth*** field, "8" in the ***Starting curve offset*** field, and "454" in the ***Ending curve offset*** field (those are the files listing numbers, as shown before every filename in the top-left panel; the names of the files selected with the Starting and Ending curve offsets are shown to the right side of each field). The selection is done by either pressing ***Select Only***, or the ***Select Additionally*** buttons (the latter will accumulate the new selections to anything previously selected in the top-right panel).
Once the selection is operated, the selected files will be highlighted in the top-side left panel, and will be listed in the right panel. The selection is then carried out to the main ***HPLC-SAXS*** window by pressing the ***Transfer selections to main window*** button.
A similar procedure operates in the ***Select by name*** subpanel.

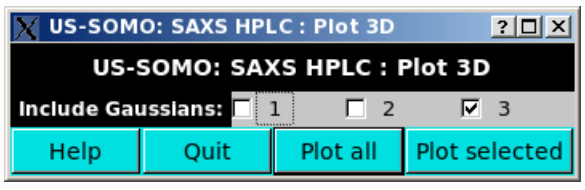
---

www contact: Emre Brookes

Last modified on March 30, 2016.

SOMO HPLC-SAXS Module 3D Plot Options Panel:

Last updated: October 2013



This pop-up panel will appear when the ***3D Plot*** button is pressed.

The *Include Gaussians:* checkboxes, whose number is automatically determined by the number of Gaussians used in the **HPLC-SAXS** Gaussian analysis, allow to decide which Gaussians are to be included in the 3D visualization.

Plot all will show all Gaussians, irrespective of what has been selected in the checkboxes above.

Plot selected will instead plot only the selected Gaussians.

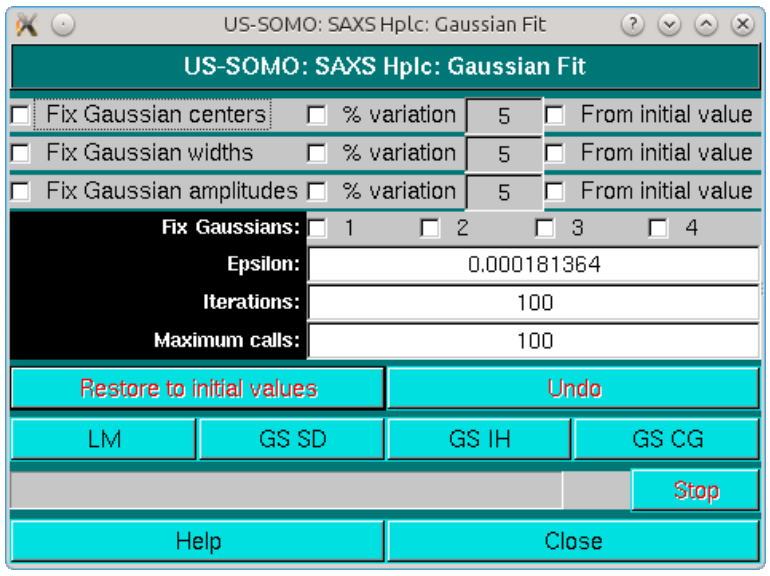
---

www contact: Emre Brookes

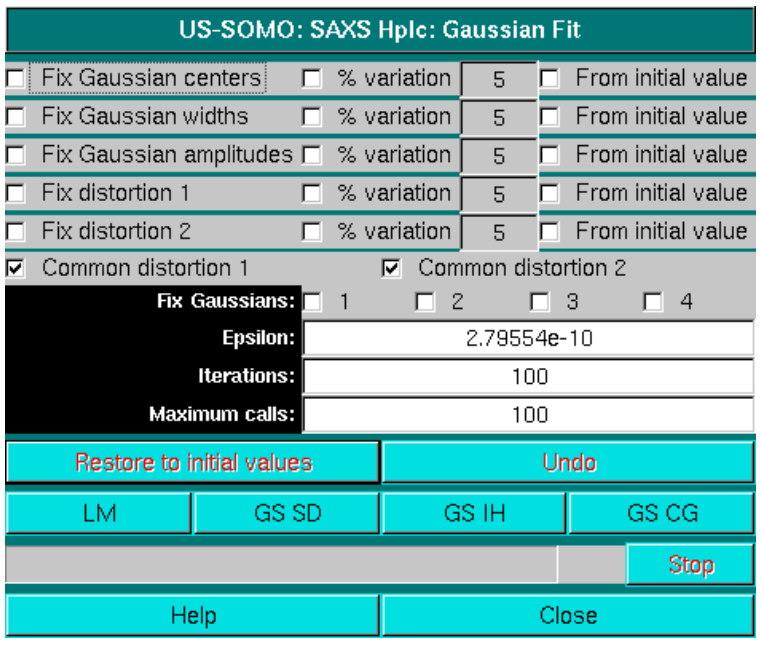
Last modified on October 31, 2013.

SOMO HPLC-SAXS Module Gaussian Fit:

Last updated: May 2014



This pop-up panel will appear when the ***Fit*** button is pressed in the main **HPLC-SAXS** module. The panel shown above refers to operations with symmetrical, non-distorted Gaussians. If distorted Gaussians are used, the panel will contain additional fields/checkboxes, as shown below for the EMG+GMG function:



The first three lines in the **Fit** panel control the centers, widths, and amplitudes of the Gaussians. For each of these parameters, it is possible to fix them to the initial values (*checkboxes* on the left side), or to allow a % variation (default: **5%**) from either the initial values (if the rightmost *checkboxes* are selected) or based on each cycle of iterations-generated values.

Importantly, since for the most common situation of different aggregation states of similar species eluting from the HPLC column one can safely assume that each peak will present a similar distorsion, by default the ***Fit*** module will present one or two (depending on the type of distorted Gaussian function) extra checkboxes called *Common distorsion #*, already selected. They can be deselected to test the effect of allowing different distorsions for every Gaussian peak. Notice that when you have fixed gaussian(s) and you have *Common distortion* set, the distortions will only be common to the non-fixed curves.

It is possible to also individually fix each Gaussian by selecting the corresponding *checkbox* on the ***Fix Gaussians*** line; the program will automatically present as many *checkboxes* as are the input Gaussians.

The **Epsilon** field controls the step used in computing the discrete derivative, gradient or Jacobian.

In the **Iterations** field the number of iterations/cycle is set (default: **100**).

The **Maximum calls** controls the attempts to improve at each stage of the minimization.

Several fitting algorithms are available through dedicated buttons: Levenberg-Marquardt (LM), Gradient Search Steepest Descent (GS-SD), Gradient Search Inverse Hessian (GS-IH), and Gradient Search Conjugate Gradient (GS-CG). When initially fitting a single chromatogram, normally a first iteration cycle is performed with the centers fixed, and then a second is sufficient with no constraint or with a restraint on the peak centers from initial values (default **5%**) to find a good set of Gaussians.
The resulting Gaussians are updated in the graphics window, and their sum is shown as a dashed yellow line, so that the goodness of the fit can also be graphically assessed. The RMSD of the fit is also updated continuously in its main panel field.

The ***Restore to initial values*** and ***Undo*** buttons are available to restart the fit procedure from the beginning, or to undo the last operation, respectively.

Close will close the **Fit** window when a satisfactory fit is obtained

---

www contact: Emre Brookes

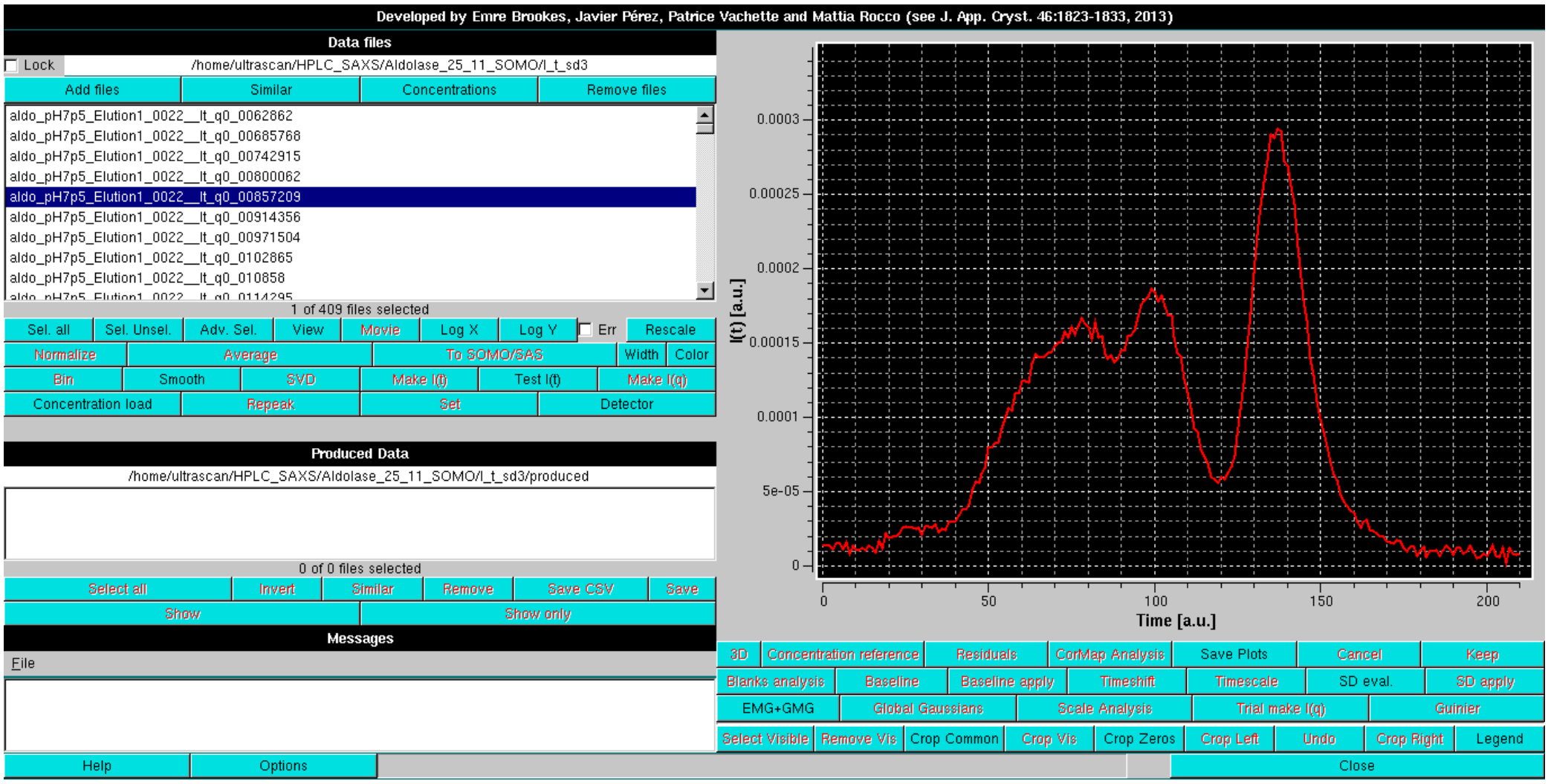
Last modified on May 15, 2014.

SOMO HPLC-SAXS Module linear baseline tool:

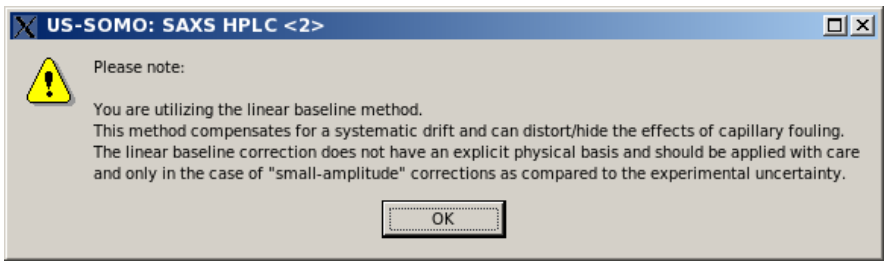
Last updated: December 2017

NOTICE: the HPLC-SAXS module is being developed by E. Brookes, J. Perez, P. Vachette, and M. Rocco.
Portions of this help file are taken from the Supplementary Materials of Brookes et al., "Fibrinogen species as resolved by HPLC-SAXS data processing within the UltraScan SOlution MOdeler (US-SOMO) enhanced SAS module", J. Appl. Cryst. 46:1823-1833 (2013), and from Brookes et al. "US-SOMO HPLC-SAXS Module: Dealing with Capillary Fouling, and Extraction of Pure Component Patterns from Poorly Resolved SEC-SAXS Data", J. Appl. Cryst. 49:1827-1841 (2016).

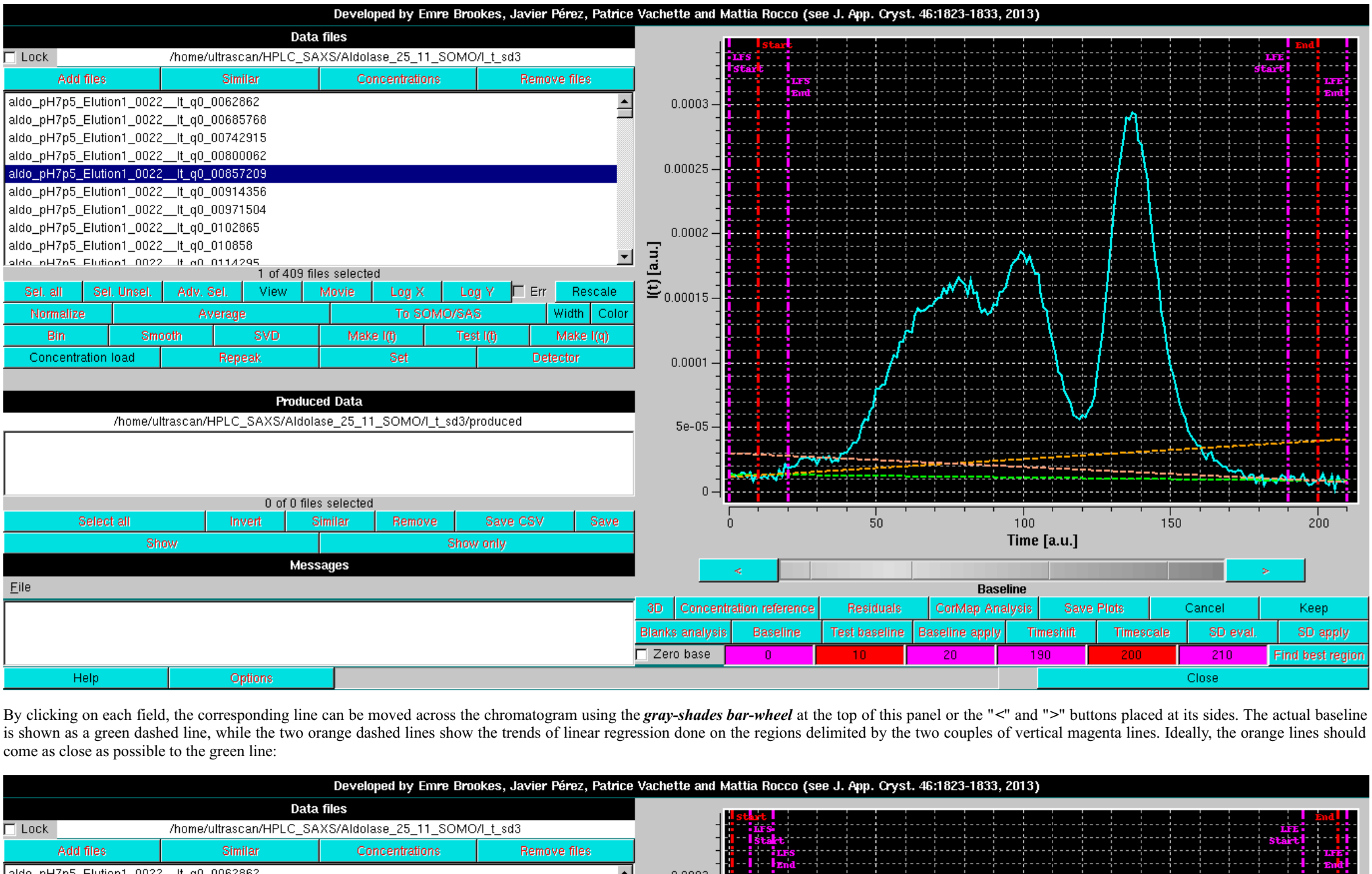
This portion of the manual describes the utilization of the original linear baseline subtraction tool (see Brookes et al. J. Appl. Cryst. 46:1823-1833, 2013). This tool should be now utilized **only** if there is just a minor drift between the initial and final baseline portions in the $I(t)$ vs. t chromatograms, likely not due to capillary fouling but to other resaons such as incomplete column buffer equilibration. Starting from the May 2016 release, we have implemented a tool assessing the amount of drift and suggesting which kind of baseline subtraction would be more appropriate.
As an example, we utilize here a HPLC-SAXS dataset collected on an Aldolase sample, which, in addition of multiple, non-resolved peaks that will be the subject of Gaussian decomposition with distorted Gaussian functions (see here), shows only a minor baseline drift:



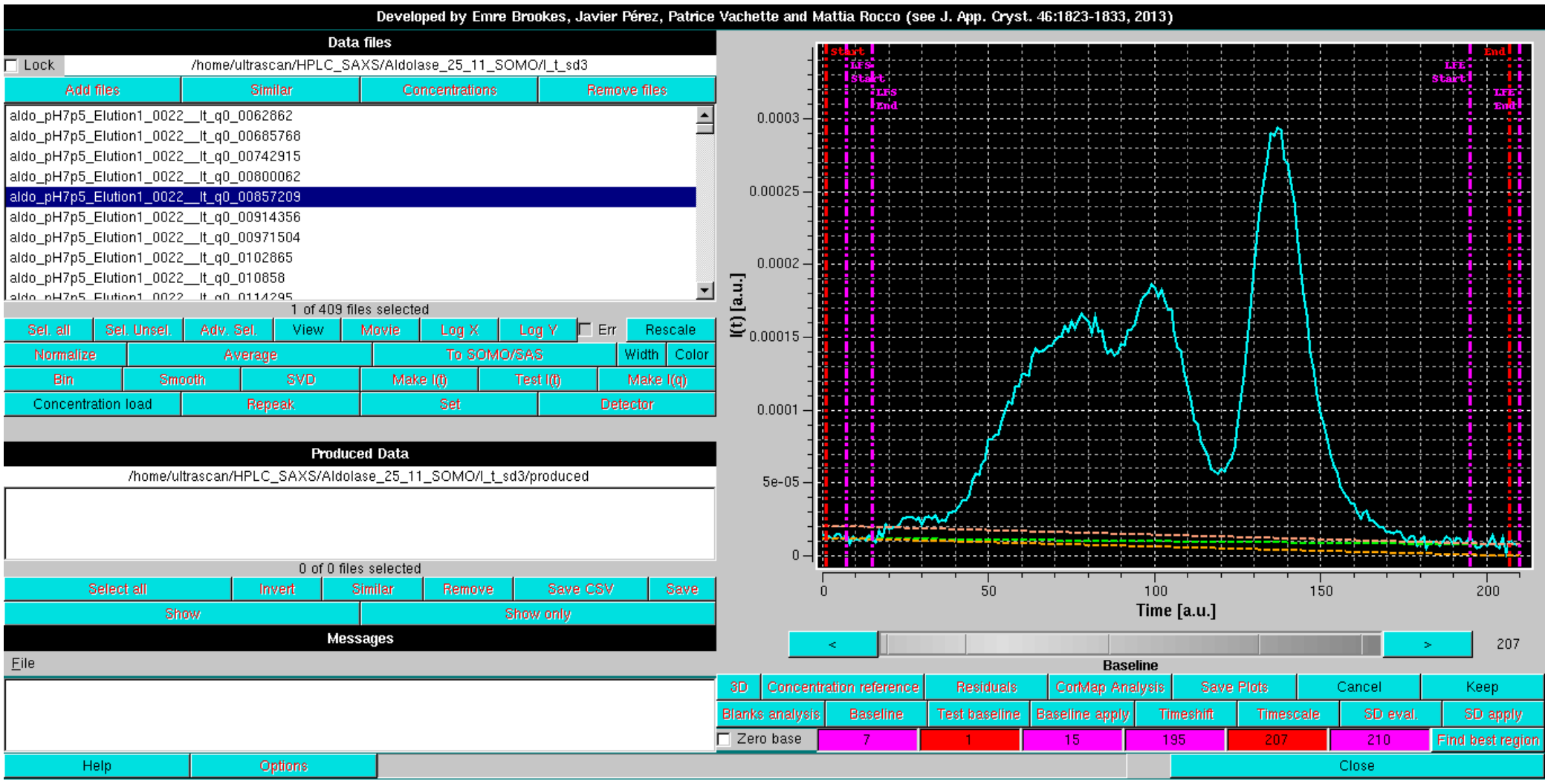
Contrary to what is required for the new Integral Baseline method, a single chromatogram is first selected; a compromise between high intensity and low noise works best.
In the ***Options*** module, select the *Linear baseline removal* checkbox. The ***Baseline*** button is then pressed. A pop-up message will appear:



Pressing "OK" will allow to proceed. As shown in the image below, this superimposes to the selected chromatogram six vertical lines, three for each side. The two magenta lines on each side define the beginning and end, respectively, of the chromatogram regions over which the data are averaged to set the beginning and end of a baseline. The red lines define instead the beginning and end points of the data to be subjected to the baseline correction. The positions of the six lines are shown in the six fields in the bottom row, with their backgrounds color-coded accordingly. If the *zero base* checkbox is selected, the left-side lines and they respective fields will be removed, as it happens for the integral baseline operation (***default: not selected***).



By clicking on each field, the corresponding line can be moved across the chromatogram using the ***gray-shades bar-wheel*** at the top of this panel or the "<" and ">" buttons placed at its sides. The actual baseline is shown as a green dashed line, while the two orange dashed lines show the trends of linear regression done on the regions delimited by the two couples of vertical magenta lines. Ideally, the orange lines should come as close as possible to the green line:

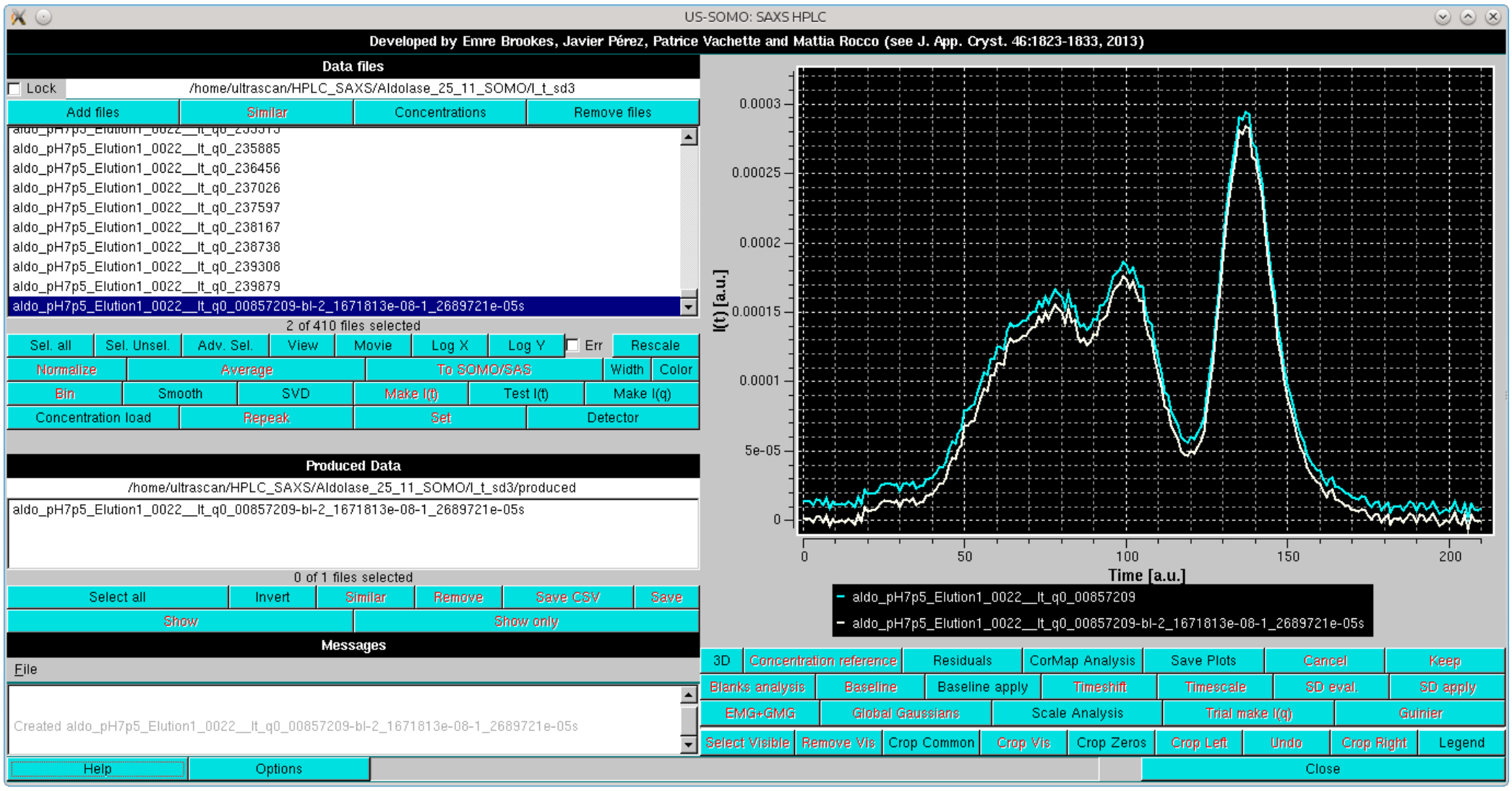


Pressing ***Keep*** once a reasonable baseline has been found will keep its parameters (initial and end points, slope) for further operations.

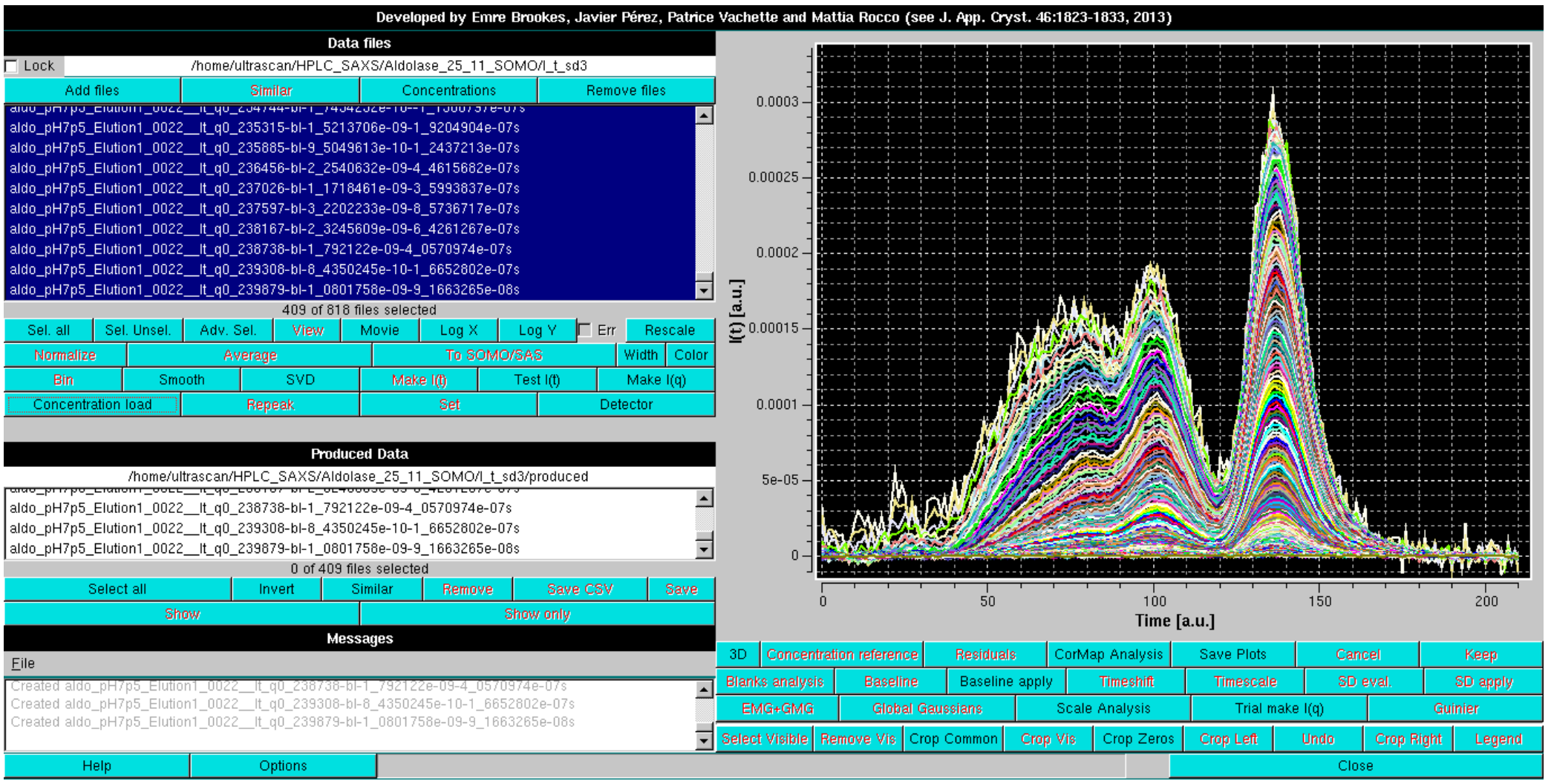
Cancel will remove the settings and revert to no baseline.

After pressing ***Keep***, it is best to then select one by one a few other chromatograms and press ***Baseline*** again to see how the chosen settings perform for other datasets. If necessary, the settings can be modified and replace the initial ones.

The ***Baseline apply*** button becomes available once a baseline has been defined even a single *I*(*t*) vs. *t* chromatograms is selected. Pressing it will allow to compare how the chosen settings perform:



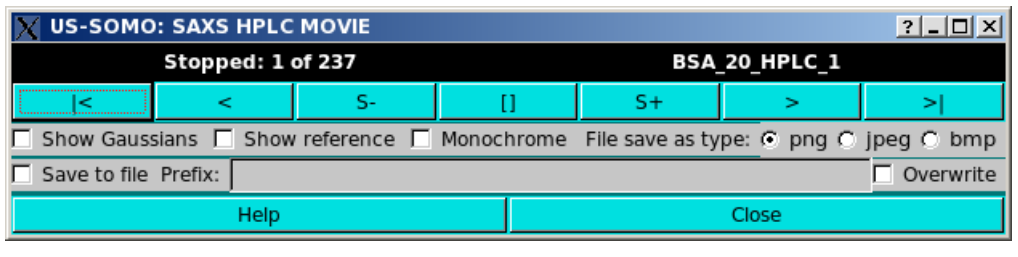
The baseline parameters thus set can be applied to all curves with concurrent subtraction of each baseline by selecting them all and then pressing again the ***Baseline apply*** button. A new set of data is generated, the initial and final points used in the linear baseline subtraction are added to the filename of the produced files, and two labels are added, "-bl" (=baseline linear) after the *q* value, ans "-s" at the end of the filename, as shown in the ***Data files*** panel:



The ***Linear Baseline***-subtracted data can then be saved and further processed as described for the ***Integral Baseline***-subtracted data.

SOMO HPLC-SAXS Module Movie Generator Utility:

Last updated: October 2013



This utility allows to view in the main graphics window of the **US-SOMO HPLC-SAXS** module a series of data files in a movie-like manner, and to optionally save each frame as an image for real movie-making operations.

In the top black box, the status of the "movie" showing is reported, with the current data file indicated.

The following box contains the controls for the "movie" operations:

- **|<** the visualization will go to the first frame (file) among those selected
- **<** the visualization will go one frame (file) back
- **S-** slows the frame rate during visualization
- **[]** will sequentially show the frames (play the "movie")
- **S+** accelerates the frame rate during visualization
- **>** the visualization will go one frame (file) forward
- **>|** the visualization will go to the last frame (file) among those selected

Below the commands line are several options:

- *Show Gaussians*: selecting this checkbox will enable the visualization of the Gaussians, if present, when *I(t)* vs. *t* frames are played
- *Show reference*: if a concentration data file (UV or refractive index monitor data) is present, selecting this checkbox will add an additional graph below the main graphics window where the concentration signal is displayed, and the time point corresponding to the actual SAXS data frame is shown as a vertical line
- *Monochrome*: if this checkbox is selected, the data frames in the main graphics panel will be shown all in a single color
- *File save as type:* ❍ *png* ❍ *jpeg* ❍ *bmp* : if the frames are to be saved as images (next checkbox), here the output image file type can be selected
- *Save to file prefix:* selecting this checkbox will enable to save each frame played as an image file, so that a real movie can then be made. A prefix can be entered in the field to the right, to be added to each frame's name
- *Overwrite*: selecting this checkbox will allow overwriting existing filenames on file saving

---

www contact: Emre Brookes

Last modified on October 18, 2013.