MetAssimulo README

Harriet Muncey

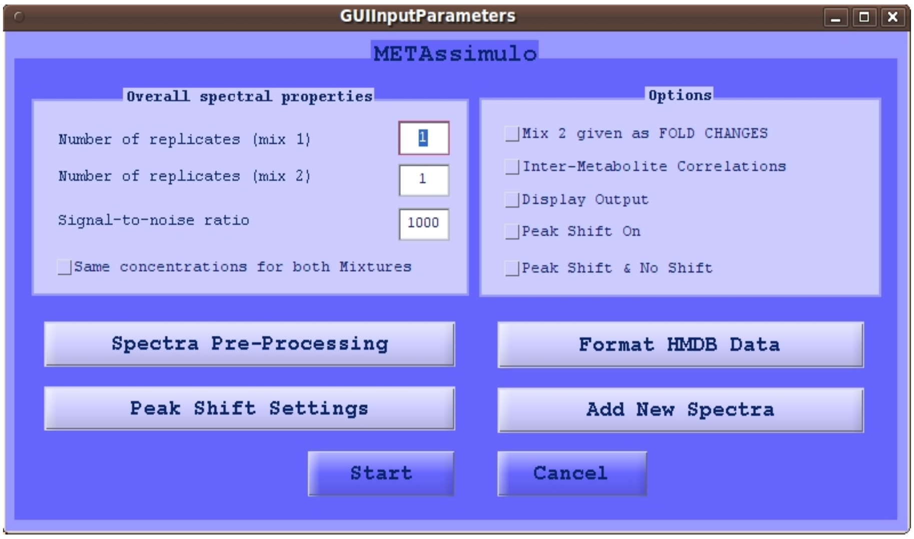
Contents

1 QUICK START

GET STARTED right away by using the default input files to simulate normal urine profile vs a paraquat poisoning profile.

DOWNLOAD the MetAssimulo bundle from the website. Included are the matlab files, the input files and the NSSD. (See Section.5 for detailed description of all input files.) Open MATLAB and ensure you are working in the MetAssimulo directory.

RUN MetAssimulo by typing 'MetAssimulo' in the MATLAB command line. The MetAssimulo GUI (Fig.1) will appear, providing access to a variety of settings. Select the option 'Mixture 2 given as fold changes' and then click 'Start'. METASSIMULO will output real-time updates on the simulation progress to the MATLAB command window, and when finished will display the mean spectra for each mixture. All simulation data for this 'run' will be output to a folder within the 'Output' directory, labelled with an appropriate time- stamp.



Overall spectral properties

Number of replicates (mix 1) 1

Number of replicates (mix 2) 1

Signal-to-noise ratio 1000

□ Same concentrations for both Mixtures

Options

□ Mix 2 given as FOLD CHANGES

□ Inter-Metabolite Correlations

□ Display Output

□ Peak Shift On

□ Peak Shift & No Shift

Figure 1: MetAssimulo GUI

2 BATCH MODE

When running large numbers of simulations, it may be useful to run MetAssimulo in 'batch mode'. To run MetAssimulo straight from the parameter file without the GUI, type 'SimulateSpectrum('batch');' in the MATLAB command line. Outputs from each run will be saved in a time-stamped folder as usual.

3 SIMULATING SPECTRA

This section describes in more detail how to simulate spectra using MetAssimulo, including a description of the outputs produced. Generation and modification of input files is described in section 5.

Type 'MetAssimulo' in the MATLAB command line to open the MetAssimulo GUI, Figure.1. Specify the number of replicates of each mixture and the desired signal to noise ratio.

You have several OPTIONS for simulation. If your second mixture concentrations are expressed as fold changes of the first, ensure you select 'Mix2 given as FOLD CHANGES'. If you wish to incorporate inter-metabolite correlations (see Section.3), again tick the appropriate box. There is also the option to produce simulations both with and without peak shift using the same simulated concentrations. The default is to simulate with peak shifts.

CLICK START to commence the simulation. The MATLAB command line will update you on the processing stage. Once finished, your simulation data will be saved in an 'Output' directory (created in the MetAssimulo directory) within a sub-folder named using a time-stamp. There are several useful output files:

- Concentrations_Mix1.txt, Concentrations_Mix2.txt - Simulated concentrations for each metabolite in each replicate of each group.

- Metabolites_NOT_Included.txt - A list of metabolites which did not contribute to the simulation (for example because their spectra or other input was unreadable).
- Correlation_and_Covariances.txt - Input correlations, the automated alterations used to ensure positive definiteness and the final covariance matrix calculated.
- pH_Mix1.txt, pH_Mix2.txt - Simulated pH values for each replicate of each group.
- pKa_list.txt - List of pKa values used (simulated and/or input).
- acid_base_list.txt - List of acid/base chemical shift values used (simulated and/or input).

If you are simulating both with and without peak shift for the same simulation, output files will be put into two folders, 'With Shift' and 'Without Shift', within the same time-stamped simulation folder. Also, a plot of the mean spectra of the two groups is output to the screen as a MATLAB figure and a jpg is saved.

4 METABOLITE CORRELATIONS

You can either load correlation data from a text file (Click on Mixture 1 or 2 in the Load Correlation Data panel of the Correlation GUI) or construct the correlation matrix using the GUI Figure.2. To edit the correlation matrix entries select the required metabolite from the drop-down menu and click 'Add'. Select the matrix cell you wish to change, enter the new value into the text box below the matrix and click the 'Update selected cell'. This method can also be used to

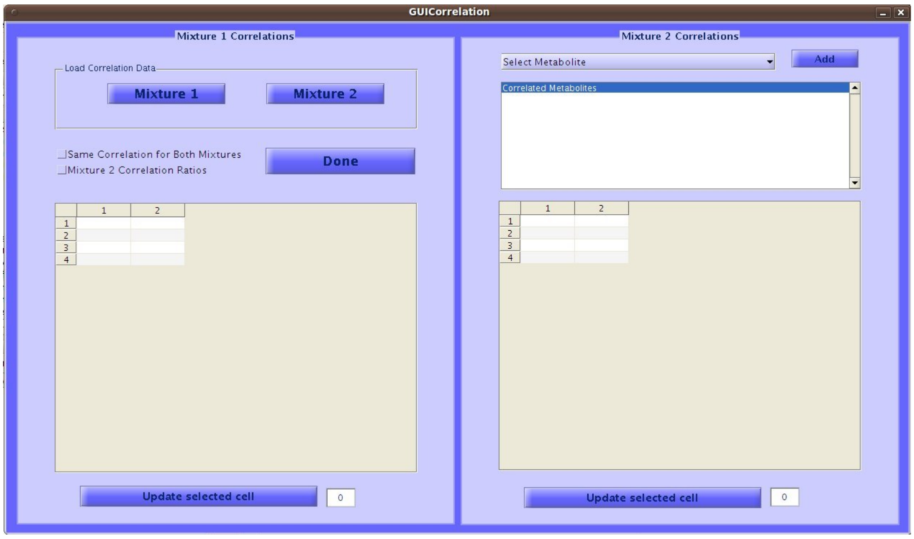


Figure 2: Correlation GUI

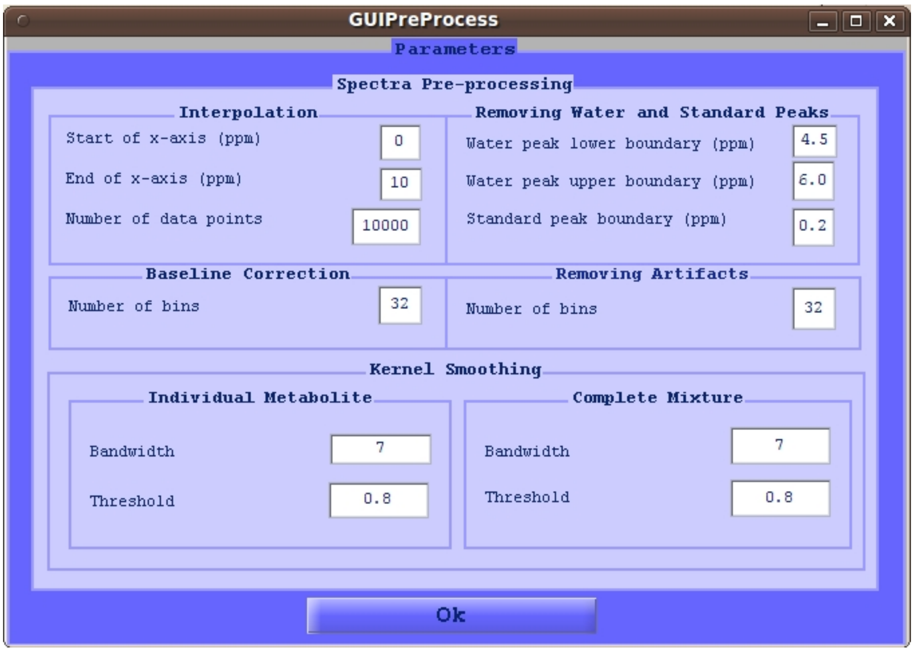
alter the entries of a matrix loaded from a text file. Simply add the metabolites whose correlations you wish to change and update the cells as before. When you have finished editing click ‘Done’.

Using the correlations you have input, MetAssimulo constructs the ‘nearest’ covariance matrix for simulating correlations. These matrices are given in the output file ‘Correlation_and_Covariances.txt’ so you can inspect the neccessary alterations.

5 ALTERING PARAMETERS

Two sets of simulation parameters can be altered via specific GUIs.

Spectra pre-processing parameters, such as exclusion region location, smoothing, and binning can be altered by clicking on ‘Spectra Pre- Processing’ in the main MetAssimulo screen.(See Figure.3).



Spectra Pre-processing

Interpolation
Start of x-axis (ppm) 0
End of x-axis (ppm) 10
Number of data points 10000

Removing Water and Standard Peaks
Water peak lower boundary (ppm) 4.5
Water peak upper boundary (ppm) 6.0
Standard peak boundary (ppm) 0.2

Baseline Correction
Number of bins 32

Removing Artifacts
Number of bins 32

Kernel Smoothing

Individual Metabolite
Bandwidth 7
Threshold 0.8

Complete Mixture
Bandwidth 7
Threshold 0.8

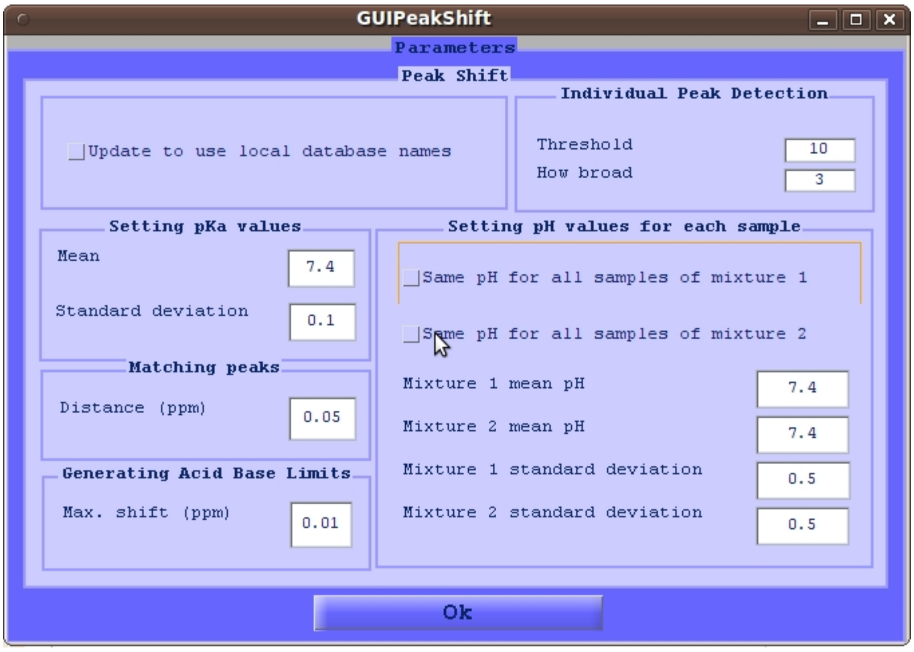
Figure 3: Spectra Pre-Processing GUI

Peak shift parameters, such as pH values and peak detection thresholds, can be adjusted by clicking on ‘Peak Shift Settings’ (See Figure.4). When calculating peak shift, MetAssimulo assumes NSSD spectra are acquired at pH 7.4.

Alternatively, all parameters can be altered in the parameters.txt file.

6 SETTING UP YOUR OWN NSSD

This section describes how to generate the input files required by MetAssimulo and their format. Most input files can be generated automatically using information downloaded from the Human Metabolome Database (HMDB). The



□ Update to use local database names

Individual Peak Detection

Threshold 10

How broad 3

Setting pKa values

Mean 7.4

Standard deviation 0.1

Matching peaks

Distance (ppm) 0.05

Generating Acid Base Limits

Max. shift (ppm) 0.01

Setting pH values for each sample

□ Same pH for all samples of mixture 1

□ Same pH for all samples of mixture 2

Mixture 1 mean pH 7.4

Mixture 2 mean pH 7.4

Mixture 1 standard deviation 0.5

Mixture 2 standard deviation 0.5

Figure 4: Peak Shift Settings GUI

information generated includes metabolite peak positions, metabolite synonyms and the normal human urine metabolite concentrations.

First, DOWNLOAD the following files from http://www.hmdb.ca/downloads and unzip. 'NMR Spectra Peaklist Files' : a folder containing peaklist data. 'MetaboCard Flat Files' : a text file of 'Metabocards'.

MetAssimulo assumes the NSSD folders are identified by the NSSD metabolite name. The NSSD directory structure and files should be in the Bruker format. A text file is needed listing each local NMR Standard Spectra Database (NSSD) metabolite name and the experiment number to be used, separated by a TAB.

e.g:
2_6-DTBut-4-MePhe 7
2_deoxyadenosine 7
etc...

RUN MetAssimulo by typing 'MetAssimulo' in the MATLAB command line. Ensure you are working in the MetAssimulo directory. Click 'Format HMDB Data'. You will then be prompted to choose the age range and sex of urine sample concentrations to be used, as well as specifying the locations of the files listed above (see Fig.5).

This function which will produce data needed for conversion between the NSSD and HMDB data, peak location lists required for simulations, and concentration data for a sample of 'normal' urine to use as a control.

OUTPUT FILES are saved in a 'Setup' folder created in the MetAssimulo directory.

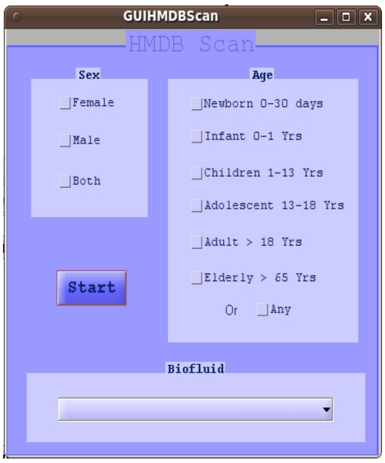


Figure 5: HMDB Scan GUI

multiplets.txt - No headerline. Each line is of the form:
HMDB metabolite name TAB peak midpoint TAB peak start TAB peak end
e.g:
Betaine s 3.89 3.84 3.93
Betaine s 3.25 3.19 3.31
Betaine s 3.25 3.19 3.31
L-ascorbic_acid d 4.51 4.48 4.53
L-ascorbic_acid ddd 4.01 3.97 4.04
L-ascorbic_acid m 3.73 3.69 3.79
L-ascorbic_acid m 3.73 3.69 3.79

peaks.txt - With headerline. Subsequent line format:
HMDB metabolite name TAB peak midpoint TAB peak intensity
e.g:
HMDB_metabolite_name Peak_position(ppm) Peak_Intensity
2_hydroxybutyrate 0.874 0.4592
2_hydroxybutyrate 0.886 1
2_hydroxybutyrate 0.899 0.4888
2_hydroxybutyrate 1.617 0.045

experiments.txt - No headerline. Each line is of the form:
”NSSD metabolite name” TAB experiment number
e.g:
”2_6-DTBut-4-MePhe” 7
”2_deoxyadenosine” 7
”2_deoxycytidine” 7
”2_hydroxybutyrate” 13
”2_NH2_5_OH_BA” 7

synonym_converter.txt - No headerline. Each line is of the form:
”NSSD metabolite name” TAB ”HMDB synonym”
e.g:

”1-Methylhistidine” ”1 methylhistidine”
”1-Methylhistidine” ”1-methyl histidine”
”1-Methylhistidine” ”1-MHis”
”1-Methylhistidine” ”1-Methyl-Histidine”
”1-Methylhistidine” ”1-Methyl-L-histidine”
”1-Methylhistidine” ”1-N-Methyl-L-histidine”
”1-Methylhistidine” ”L-1-Methylhistidine”
”1-Methylhistidine” ”N1-Methyl-L-histidine”
”1-Methylhistidine” ”pi-methylhistidine”

local_names.txt - No headerline. Each line is of the form:
NSSD metabolite name TAB HMDB metabolite name
e.g:
2_oxobutyrate 2-Ketobutyric acid
2_hydroxybutyrate 2-Hydroxybutyric acid
L-carnosine Carnosine
n-butyrate Butyric acid

COPY THESE FILES to your MetAssimulo ‘Input’ directory. ‘multiplets.txt’ and ‘peaks.txt’ should be put in a subfolder ‘peak_shift’. If you have pKa values and acid/base limits available, this data also belongs in the ‘peak_shift’ folder (If this data is unavailable, values will be simulated in MetAssimulo). You should use two separate files with the following formats.
pKa_list.txt - No headerline. Each line is of the format:
HMDB metabolite name TAB pKa value
e.g:
1-Aminno-2,2-dimethylpropane 10.15
1-Amino-6-hydroxynaphthalene 3.97
1-Methylimidazol 6.95
1-Methylpiperidine 10.08
1-Naphthoic acid 3.7

acid_base_list.txt - No headerline. Each line is of the format: HMDB metabolite name TAB peak midpoint TAB acid limit TAB base limit
e.g:
Creatinine 4.08429 4.06533 4.08733
Glycine 3.56536 3.55892 3.56916
Citric_acid 2.55726 2.55706 2.55774

OTHER OUTPUTS are two concentration files ‘raw_mmolcr_concentrations.txt’, giving concentrations extracted from the HMDB whose units are umol/mmol of creatinine and ‘raw_uM_concentrations.txt’ whose units are uM. These are given in the format required for inputting concentration data to MetAssimulo. MetAssimulo simulates two groups of spectra so your concentrations for both must also be in this format:
normal_urine.txt - With headerline. Subsequent line format:
”NSSD metabolite name” TAB mean TAB standard deviation
e.g:
NMR STANDARDS MEAN ST DEV
”Creatinine” 13200 4100

”Citric acid” 2022 1081
”Glycine” 1029 440
”L-Histidine” 948 371

MISSING DATA is reported in ‘could_not_find_data.txt’. It lists metabolites without matching HMDB names, peak data or concentrations. This parse of the available data should pick up most but probably not all the information required, due to unpredictable human errors in compiling the HMDB. It is suggested the user checks through these to see if any NSSD names can be modified to find a match in the HMDB and manually alters files as needed.

ALSO REQUIRED is ‘protons.txt’: a file containing the number of protons for each metabolite must be specified.
protons.txt - No headerline. Each line is of the form:
”NSSD metabolite name” TAB proton number
e.g:
”3-methyluridine” 11
”adenosine” 8
”Arabitol” 7
”L-ascorbic_acid” 4
”Phenylacetic_acid” 7

ONCE YOU HAVE COMPLETED THESE STEPS you should have 5 files and a subfolder within your ‘Input’ directory:

- normal_urine.txt (either from the outputted concentrations or your own data)
- protons.txt
- experiments.txt
- local_names.txt
- synonym_converter.txt
- peak_shift (folder)
    - peaks.txt
    - multiplets.txt
    - acid_base_limits.txt (optional)
    - pKa_list.txt (optional)

Ensure that ’parameters.txt’ correctly identifies the location of your NSSD. When you run your simulations in MetAssimulo, if the peak shift input files are not formatted to the NSSD names (i.e. you have used peak data from elsewhere), you must ensure you select ‘Update to Local Database Names’ by clicking on ‘Peak Shift Settings’. This will output the peak shift data with the appropriate metabolite names which can then be used as input so as to save computational effort: converted_multiplets.txt, converted_peaks.txt, converted_pKa_list.txt, converted_acid_base_list.txt - The same as standard peak shift input files, except all metabolite names are converted to the NSSD metabolite names.

NB - MetAssimulo assumes the NSSD folders are identified by the NSSD metabolite name.

7 ADDING NEW SPECTRA

Copy your spectra files to your NSSD ensuring the files are structure in the Bruker format. Run MetAssimulo and click the ‘Add New Spectra’ button. Browse for the metabolite you wish to add and select the folder of the experiment number you wish to use (see Fig.6). It is possible to add multiple metabolite spectra simulataneously by browsing for each folder and adding it to the list displayed.

Click ‘Start’ to begin. You must specify the location of the ‘metabocards_all.txt’ file, the HMDB NMR Peaklist folder and the input files you wish to update. Then, all required input for the new metabolites will be extracted from the HMDB data and appended to your input files without altering existing content. ‘could_not_find_data.txt’ will report any problems searching for the HMDB data. ‘protons.txt’ and your normal urine template will need to be updated manually to include the new metabolite.

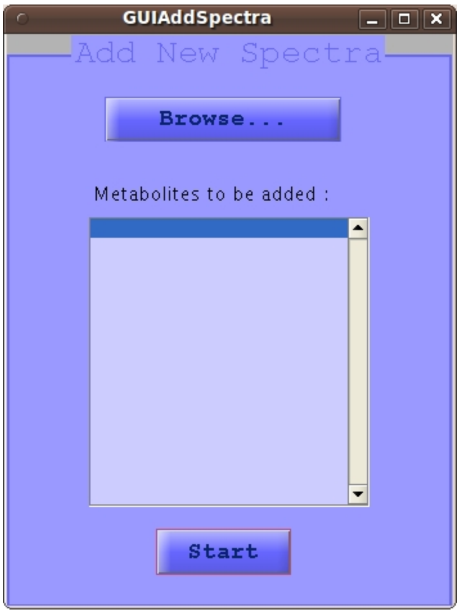


Figure 6: Add New Spectra GUI

8 SPECTRA PRE-PROCESSING DETAILS

This section describes the preprocessing applied to the pure compound spectra in more detail. This should help the user understand the effect of altering different parameters.

8.1 Exclusion Regions

Note that for biofluids other than urine the upper boundary can be shifted (e.g. from the default 6.0ppm to 5.0ppm). This is important as there are some metabolites which have resonances within this region, e.g. glucose, which obviously will not be included with the default exclusion settings.

8.2 Baseline Correction

It is easier to distinguish peaks in a spectrum when the baseline is featureless, however, spectra can have distorted baselines due to imperfections in the detection process. Curved baselines can be a major source of error and so a correction is carried out on the spectrum using a moving average. This method involves splitting the data into smaller 'windows' of data then using the average within the window to smooth out the spectrum. In order to alter the baseline without losing the resolution of the peaks, a threshold is set by dividing the maximum peak by a user specified parameter. Then all the intensities found below this threshold are corrected.

The moving average window size, ω, is defined by the user. ω_{mid} is the window mid point (corresponding to the data point, not the chemical shift). The window starts at y_1 and moves along all the data points in the spectrum until it reaches the final point, y_n. The median of the values in the each window is calculated and then used to generate an interpolating spline over the number of points in the whole spectrum. This is then then subtracted point by point from the spectrum.

It is important to note that the baseline correction often produces negative artifacts due to the fact that it is difficult to smooth only the baseline. Positive artifacts may also appear if the input spectra contain a wide urea/water peak that affect the baseline outside the exclusion region. This can be problematic as a build up can lead to false peaks in the mixture spectrum. A simple way to remedy this problem is to alter the exclusion region parameters.

8.3 Removal of Negative Artifacts

Negative artifacts, produced by baseline correction or simply inherent in the original spectrum must be removed since their presence could interfere with peaks of interest in the mixture spectrum. This is remedied by using an estimate of the noise standard deviation, σ_{med}, to calculate a limit value, l, using the equation below. σ_{med} is estimated by splitting the spectrum into a number of bins (given by the user, default 32) and calculating their standard deviations. The median of these standard deviations is used as the estimate of σ_{med}.

$$l = M - 3\sigma_{med}$$

M is the median of the intensities, y_i. All intensities appearing below this limit, l, are set equal to it.

8.4 Kernel Smoothing

When adding the spectra of the metabolites together to form the final spectrum, noise from each will also be combined, reducing the overall signal to noise ratio. Kernel smoothing is used to reduce the noise in each individual metabolite spectrum. This process estimates the 'smooth' function underlying the noisy set of data points by looking at each individual point and measuring the influence of the surrounding data points according to the choice of kernel used. Whilst the default kernel type is 'Normal', the user may instead specify Cauchy, Laplace, Uniform, Epanechnikov, Biweight, Triweight, Triangular, or Logistic in the parameter file. The user can also alter the bandwidth (as a number of data points), which controls how wide the probability mass is spread about the individual points. In order to minimise the risk of reducing the resolution of the peaks, only intensities below a user-defined threshold (a percentage of the maximum intensity) are subject to the kernel smoothing.

9 TROUBLESHOOTING

- Ensure input files are formatted correctly (check against provided input).
- Ensure filepaths names are correct (check parameter.txt and update as required). Ensure NSSD names correctly identify spectra folders and that the NSSD is in the Bruker format.
- Ensure proton number exist and is non-zero, otherwise the default is 1 and the metabolite will appear, but at an altered level.
- Check correspondence between HMDB peak locations and actual peak locations in experimental data. Individual metabolite template spectra are available at `http://www.hmdb.ca/downloads` if required.
- If possible, try to avoid spectra whose peaks appear only in the exclusion region or are very noisy.