

# Metabolomics

AnalyzerPro<sup>®</sup> in your Metabolomics Workflow



The simple solution to unravel complex systems in targeted and untargeted metabolomics workflows

**SpectralWorks**



## What are the Benefits of AnalyzerPro?

AnalyzerPro is vendor neutral and can convert any vendor's profile or centroid data format quickly. With fast data file processing time for noise reduction, peak picking and deconvolution, coupled with automated RI calculation (chromatographic alignment), internal standard normalization, library generation based on components in your samples and library searching (e.g. NIST/EPA/NIH mass spectral libraries), AnalyzerPro offers a complete pre- and post- processing solution to your untargeted and/or targeted metabolomics workflows through speed and automation.

AnalyzerPro has a user friendly, intuitive, interface with a convenient, adjustable sample display to show the progress of your data processing sequence. With access to all information throughout processing, including the ability to view chromatograms, raw mass spectra and deconvoluted spectra, AnalyzerPro offers a processing workflow suitable for beginners and advanced MS acquisition experiments.

## The Metabolomics Workflow

### Research Question

The most important part of any metabolomics study is the specific biological question that the study is designed to answer. This question should be used to define the overall study and should be objective, measurable, realistic and relevant. For example: is there a difference in the secreted metabolome of Salmonella grown as a biofilm or in a planktonic form? Or: which metabolites are increased in the bloodstream following inflammation? Without this biological question it may be impossible to make sense of the data or to design the study appropriately. It is not sufficient to simply collect as much data as possible and then attempt to explain observations.

### Experimental Design

It may seem obvious, but the study should be designed to answer the biological question. As we are dealing with biological processes it is critical that we have enough samples in the study to account for natural biological variation. The preference is for as many biological replicates as possible per treatment and control. The metabolomics Society recommends five replicates and the minimum number required for publication is three. The actual number of samples required for a given study will vary depending on a number of factors. For example, human studies can be particularly problematic due to the variation in human subjects and as such require the greatest number of samples.

### AnalyzerPro at a Glance

- Deconvolution of GC-MS & LC-MS data
- Supports all major manufacturer's instruments and legacy instruments
- Targeted and untargeted analysis
- Extraction of closely co-eluting peaks
- Data alignment
- Quantitation
- Statistical analysis, PCA and ANOVA
- Elemental composition
- Integration with RemoteAnalyzer
- Seamless Integration with NIST libraries
- Accurate mass compatible
- Background file subtraction
- Tabular display of results
- Custom reports

Experimental design also includes how the MS data is going to be acquired. Enough sample must be collected to allow for pooled biological replicates. Preferably, sampling should be complete before acquisition commences so that acquisition order can be fully randomized or suitably stratified. It doesn't hurt to randomize the order samples are prepared either if it is expected to take a long time. The pooled QC samples are important not only for instrument equilibration and to correct for drift but for processing using AnalyzerPro, too. Pooled QCs are useful for target component library generation, RI calculation, and metabolite identification as they are essentially a representation of your entire sample set.

In your MS experimental design, include sample collection and preparation blanks so that artefacts of sample preparation can be eliminated from your final feature list.

### Sample Collection

Optimal sample collection methods are not always achievable; nevertheless samples must be collected in a consistent manner. If the samples contain cells which are still metabolically active, this metabolism must be quenched usually by snap-freezing or with cold solvents. Samples should be frozen and kept at as low a temperature as possible. However there is no consensus as to how long samples remain viable for metabolomics studies even when frozen at  $-80^{\circ}\text{C}$ . In general, metabolomics studies are more time sensitive than the other 'omics' (proteomics and genomics).

### Sample Preparation

The sample preparation required will vary depending on the analytical technique(s) chosen. Most methods require an extraction step to remove proteins and, if necessary, break open intact cells. The extraction stage can also be used to concentrate the analytes of interest and introduce a suitable internal standard. Analysis by GC-MS requires that the analytes are volatile and this may require the preparation of suitable derivatives. Sample preparation should also consider the requirements for long term sample storage, not forgetting that samples are usually best kept as frozen dried extracts.

### Acquisition

It is important to consider how long samples are likely to remain viable as well as how frequently the analytical instrument requires maintenance and calibration when planning the frequency and positioning of replicates. Depending on the design of the metabolomic study, the MS data may be acquired in full scan, SIM or MRM mode. Data from multiple vendors can be imported into AnalyzerPro for data processing. RemoteAnalyzer<sup>®</sup> can be used to simplify and unify the interface for submitting samples for acquisition particularly when multiple vendors' instruments are available.

### Data File Conversion, Deconvolution and Feature Detection

Data from the analytical instrument needs to be processed before it can be analysed and interpreted. Our first step in data processing is called 'deconvolution' which is defined as 'reversing the effects of convolution'. For chromatographic data the biggest problem is co-elution of chromatographic peaks. Many components will not be completely separated by the column. Using AnalyzerPro we can determine which masses belong to which of the co-eluting peaks. The deconvolution step also allows the creation of target component lists. For comprehensive data analysis automated library building is available in AnalyzerPro which includes all the components within an entire data set. This differs from a traditional target component list generated from standards in that the components do not need to be identified until later in the analysis. At this stage, it is only important that they are there. The list produced from the automated library building stage will be comprehensive and the component list will be very large. As such, there is a fundamental need for a detailed level of quality control. The upside of this is that your data set can be completely defined by the component list generated during the automated library building stage. Following deconvolution, each component found can be assigned an identity. This identity does not have to be absolute and subsequent analysis can be performed on identified components or by treating each component as an unknown. Components can be identified from libraries or from a specific target component library. A target component library is usually based on standards which have been run on in house instruments as this provides more confident identification. Absolute identity of all components is not required at this stage

but where a component of interest is found (i.e. where a difference is observed) that component should be investigated further to confirm its identity.

## Statistical Analysis

The most common way to make sense of large metabolomics data sets is by the use of unsupervised multivariate statistical analysis. Principal Component Analysis (PCA) allows the data to be modelled and to determine which metabolites [components] are contributing most to the variance that has been observed. It also allows the identification of outliers that are likely to contribute to the skewing of the variance of a particular component. PCA points you in the direction needed for interpretation, but is only one part of the data analysis workflow. The next step is to scrutinize those components contributing most to the variance. It is important to differentiate between components that may be artefacts or interferences which are of no biological interest (which can legitimately be removed or their variance accounted for) and those that are identified as causing variation in the data sets that require further investigation.

## Interpretation of Results

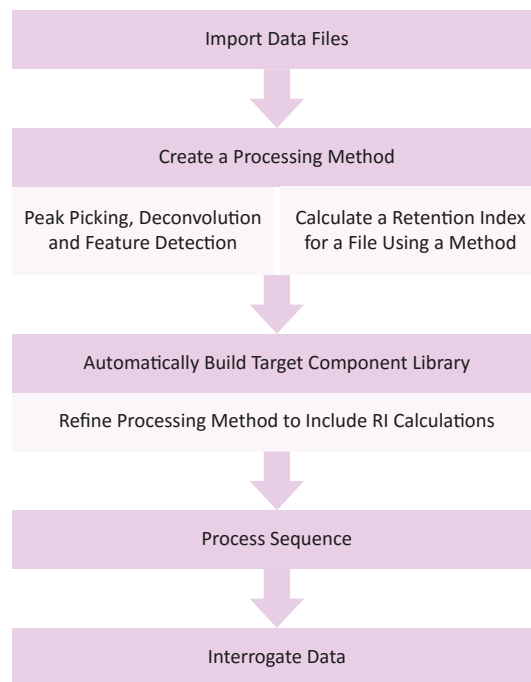
At this point in the workflow, can we answer the biological question that was posed at the beginning of the study? Is there now a new question? In either case we have moved forward in our understanding, following this metabolomics workflow. Subsequent data analysis can involve the mapping of metabolites to pathways and determining trends of specific metabolites. However, we may have indeed found that we have answered our original question. Additionally, we may have also generated many other questions. In this situation, the process of metabolomics research begins again, starting with a new biological question and if necessary, revisiting the original data set. These stages are customary practice in the analysis of metabolomic data sets.

You may have identified a biomarker or handful of biomarkers relevant to your research question that now require validation. If the goal is to use these markers for quantitation and subsequent screening in populations, AnalyzerPro can be used for quantitative data processing and can do qualitative and quantitative processing simultaneously.

## Conclusion

AnalyzerPro is a key element in this metabolomics workflow. The targeted and untargeted approach of the peak picking and deconvolution functionality is applicable to large data sets from multiple vendors' instruments. The deconvolution of complex chromatographic separations is equally applicable to GC-MS and LC-MS analyses.

## Using AnalyzerPro in the Metabolomics Workflow



## Setting Up a Processing Method

The processing method is where you determine your peak detection parameters. Your method won't be complete until you've created an RI calculation file (if required) and created a library to match against but you will need to process a data file or two to set these up so follow the instructions below to process a file and then follow **Calculate a Retention Index** for each file using a method and **Automatically Build Target Library** to complete your untargeted metabolomics data workflow method.

To create a new method, select **Method | New** and then to edit select **Method | Edit**.

Define processing constraints such as the **Minimum Masses** required to define a peak, the **Mass Range** to process, the retention time (**RT**) window to process and any masses to reject (**Reject Mass(es)**).

Define the detection parameters. This includes **Area Threshold**, **Height Threshold**, **Width Threshold**, **Signal to Noise**, **Scan Windows**, chromatographic **Resolution** and Gaussian **Smoothing**.

If you have accurate mass data, check the **Accurate Mass Processing** check box and enter the actual **Precision** (ppm) of your data and select the **Linearity** of that precision. You also need to set the precision in your preferences so that displayed spectra reports m/z to the required decimal places.

Select the **Enable target component searching?** check box. Any one of the available target component libraries can be selected from the list box. Found components will be searched against the selected library and the results will be available in the **Target Component Report**. Select the **Enable library searching?** check box if you want to enable this function. Found components will be searched against the selected library or libraries such as NIST.

Select the type, if any, of retention index processing that is required from the drop down list. Retention indices is covered in more detail in the following section.

To save these settings to the method file select **Method | Save** or **Method | Save As**. Unless these changes are saved to a method file, you may lose them.

### Calculate a Retention Index

Retention indices such as Kovats for organic compounds are a standard used to calculate a 'retention index' for a compound relative to, in the case of Kovats, hydrocarbons. For example, n-dodecane (C12) has a retention index of 1200 and C15 has a retention index of 1500. Compounds eluting between C12-15 therefore have retention indices of 1200-1500. Retention indices are useful because they are constant regardless of

temperature program and column properties such as length and diameter so compounds analysed under different conditions can be matched to a library using RI rather than retention time.

There are three ways in AnalyzerPro to calculate a RI ladder..

1. Manually
2. Using a method and a single data file and
3. Using a method on each data file.

In metabolomics experiments, especially LC-MS and large-scale GC-MS (where column maintenance is performed) where retention time shifts occur, using the 'calculate a retention index for each file using a method' is the best option.

Choose an appropriate data file to build your RI method (.rim file). This can be any file from the experiment acquisition which contains the retention index compounds. A mid-run pooled QC sample would be a good choice if available. Process the file using appropriate peak detection parameters and for speed, turn off library searching.

Once the file is processed find your retention index compounds and record their retention times. For ease, extract masses of interest (right click chromatogram and choose **Options | Trace Type | Mass Chromatogram**) i.e. for n-alkanes extract mass 57, 71, 85 etc. Then select **Home | Results | Retention Index** to create your method (.rim file)

Delete components you do not wish to use for RI calculation. Using retention time, identify your RI compounds and give them the appropriate name and retention index i.e. C12 or n-Dodecane; RI 1200. You will need to choose a retention time window based on your data. For large metabolomics studies, there is the possibility of RT shift/drift even for GC-MS. Try to quantify the RT shift before choosing the window. You will also need to include a relative intensity window. By selecting relative intensity of 50%, only those ions which have a relative intensity of 50% or greater will be used for matching. If there is any co elution or difficulty deconvoluting RI components, by using a relative intensity of 50%, ions which are 'impurities' will not be included in the matching process allowing a better match with each file.

It's a good idea to re-analyze the sample or a couple of samples

to make sure that your RIs are correct. Since you may like to use a method with RI to match spectra rather than RT, it's important to get it right or you won't get any library hits. To do this, view the summary report by selecting **Report | Summary**.

Now you can edit your processing method to include your retention index calculation method.

### Automatically Build Target Component Library

Your data set can be completely defined by the component library generated during the automated library building stage. Following deconvolution, each component found can be assigned an identity. This identity does not have to be absolute and subsequent analysis can be performed on identified components or by treating each component as an unknown. Components can be identified automatically by matching to public libraries or from a specific in-house library of authentic standards. Absolute identity of all components is not required at this stage but where a component of interest is found, that component should be investigated further to confirm its identity.

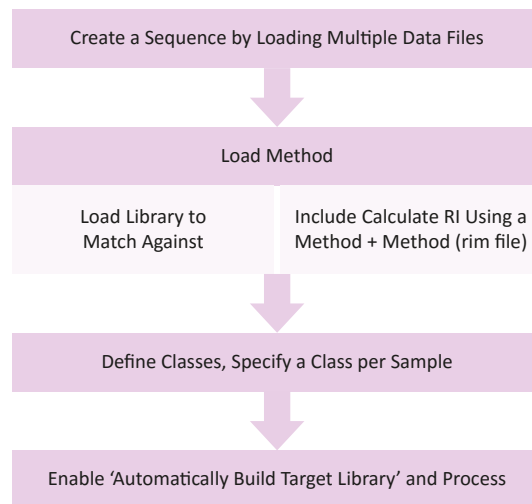
Libraries are best built using QC samples but they can be built using your entire data set, too. If you have used internal standards in your analysis, you can find them in the file you processed to create your .rim file. You can:

1. Create library from a single file
2. Automatically build library from a sequence of files

You may like to find your internal standards in your data and add these to a new library first. You can then append this library with the new components from 1) your single file or 2) your sequence. If you have QC files it's not necessary to do this step because you will need to re-process anyway but if you don't have QC files, and want to process your sequence at the same time you build your library, find your internal standard first so that you can normalize immediately.

When you are ready to build your library, go to **Home | Sequence | Analyze**. Specify the name and output location of your sequence and then check the 'automatically build target library' button.

### Processing a Sequence



### View Sequence Results

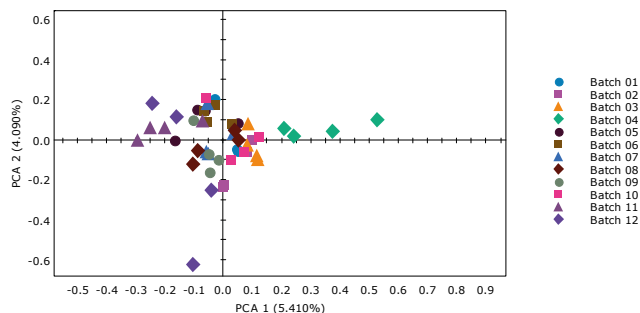
Large metabolomics data sets can be modelled using principal component analysis to visualize which metabolites [components] contribute most to the variance observed. It also allows the identification of outliers that are likely to contribute to the skewing of the variance of a particular component. PCA points you in the direction needed for interpretation, but is only one part of the data analysis workflow. Fold-changes and p-values can be calculated in AnalyzerPro to aid in determining key metabolites of interest that may help to answer the biological question.

Once your sequence has finished processing, click MatrixAnalyzer on the results section of the **Home** tab. Here you will be able to view your PCA (if desired; with or without log transformation), results and normalized results. You can edit the class information post processing to generate a PCA (**Home | Sequence | Class Information**). Press the options button (left button on the tool bar) to select which information to include in your matrix and which data transformation to use to generate your PCA.

As you scroll through components, you can inspect RT shift in your data and any main class effects for further interpretation.

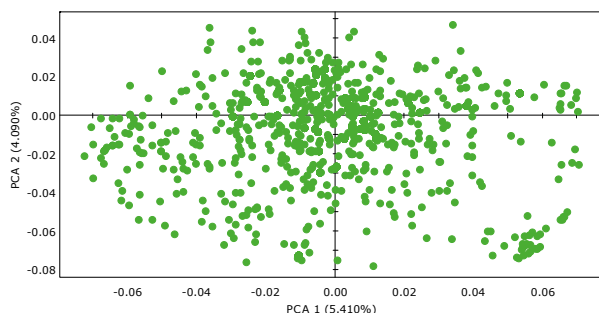
As an example, the below PCA represents 12 derivatization batches for a mixture of metabolite standards. The PCA shows that batch 4 is clearly distinct from the other batches but with a total explained variance of 9.5% for PCs 1 and 2, the batches appear fairly similar.

PCA Scores from results with log transformation



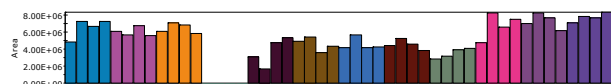
If we investigate the loadings plot, components on the left will have a higher relative concentration in batches 11 and 12 and components on the right will be higher in batch 4.

PCA Loadings from results with log transformation



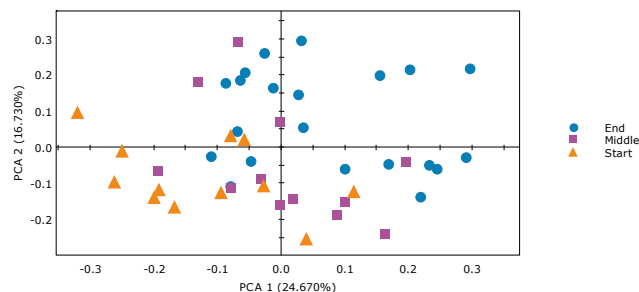
By clicking on points in the loadings plot, the bar graph, which is color coded by class, will update to show the relative concentration of the component in each sample. Below we can see that component 'QC9\_030\_22' is missing from batch 4 which is contributing to the separation of batch 4 from the rest of the QC samples.

Areas for component QC9\_030\_22

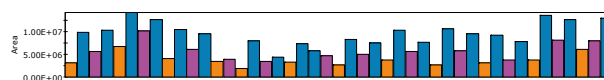


Using the same data and re-classifying the samples to their run order and removing (by selecting the row trash can button) features which occur in less than 50% of the data set, we can see a trend based on where in the analysis the sample was run. By using the scores plot and investigating the components, it can be seen that components like tentatively identified 2-piperidinecarboxylic acid have a higher relative concentration by the end of the derivatization batch.

PCA Scores from results



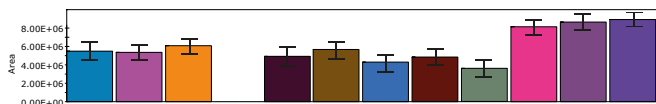
Areas for component 2-Piperidinecarboxylic acid, 1-(trimethylsilyl)-, trimethylsilyl ester



If you defined classes in your sequence, fold changes can be calculated and displayed in the MatrixAnalyzer Results Viewer. The fold change value represents the average peak area response of one class over another and is a useful measure to determine changes in compounds as a result of an experimental condition. The MatrixAnalyzer results can be sorted according to fold change.

By specifying your classes, you will also be able to view each component by experimental group plotted as the mean  $\pm$  standard deviation. Group means are color coded consistent with the PCA plot (according to class). This representation allows clear visualization of a components' relative concentration, further aiding the user's interpretation of results.

Average Areas per Class



You can calculate a p-value using, if appropriate, a one-way ANOVA. The p-value, used in hypothesis testing, indicates that there is a statistically significant difference between one group mean when compared to another. A significant result (usually if the p-value is less than 0.05) indicates that the probability of the null hypothesis being accepted is very low (5%) and the alternative (test) hypothesis can be accepted. As with fold change, results can be sorted by p-value.

AnalyzerPro will also generate a volcano plot to visualize all components plotted against their fold change and p-values. This representation of data allows the user to determine those components with the highest fold changes and lowest p-values (i.e. the most important results) with ease.

# ANALYZERPRO®

For a free 15 day trial licence, visit our website:

<http://www.spectralworks.com>

## SpectralWorks Limited

The Heath Business and Technical Park

Runcorn, Cheshire, WA7 4EB, United Kingdom

T: +44 (0)161 327 2143 | E: [info@spectralworks.com](mailto:info@spectralworks.com)

## About SpectralWorks Limited

SpectralWorks Limited is a leading UK based, privately owned software development company.

We are dedicated to providing innovative solutions targeted for markets within the life sciences industry and have strong working relationships with a number of the major instrument manufacturers. Coupled with our collaborations in academia and industry, we believe we have the right balance between scientific and software development expertise to provide the best scientific solutions for the end user.

Our vision at SpectralWorks is to improve the way software is integrated within the laboratory environment by providing the correct solutions to increase productivity and reduce overheads. We achieve this by maintaining focus on the end user, listening to their requirements and ensuring that they have the right tools to handle their day to day tasks.

In addition to the wide range of mass spectrometry software products, we provide highly respected consultancy services. These services can cover the complete software development cycle or specific steps within a projects' life cycle. These services are frequently utilized by instrument vendors and end users that demand the best for their laboratories.

As well as direct sales, we have partner resellers and offer Original Equipment Manufacturers (OEM) the ability to distribute our products and algorithms using their own branding. We also provide custom development for individual companies looking to automate and unify the software used within their laboratories using our extensive industry experience.



Certificate Number 13452  
ISO 9001:2008

This publication provides outline information only which (unless specifically agreed by SpectralWorks Limited in writing) may not be used, applied or reproduced for any purpose or form part of any order or contract or be regarded as a representation relating to the products or services concerned. SpectralWorks Limited reserves the rights to alter, without notice, the specification, design, price or conditions of supply of any product or services.