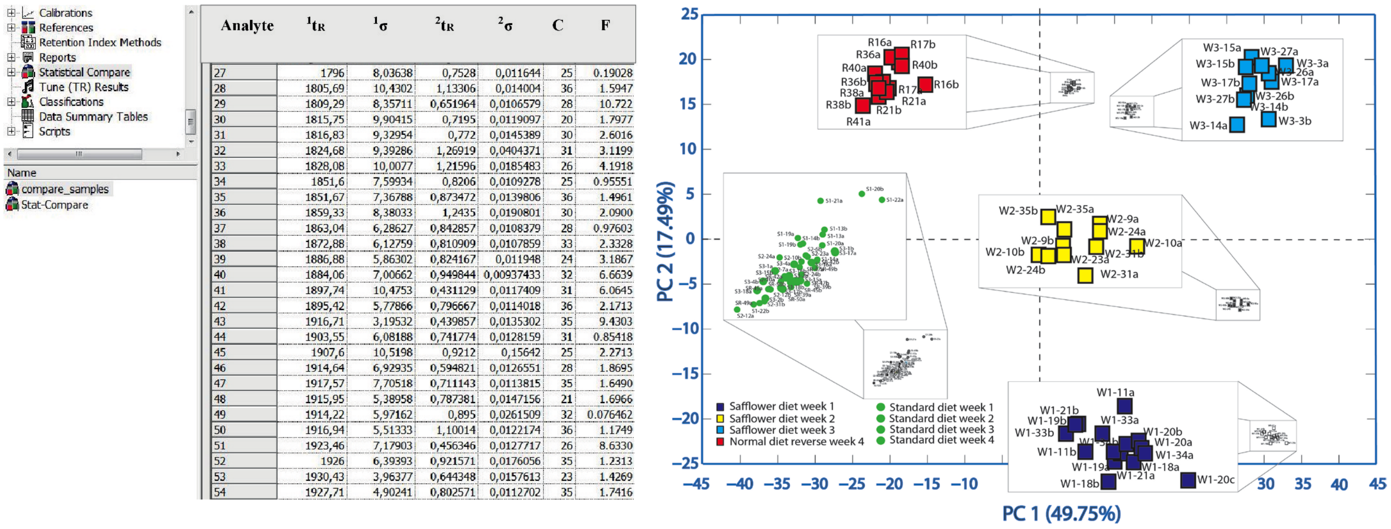
New Platform-Independent Data Analysis Software with Build-in Chemometric Tools for the Processing and Statistical Analysis of Comprehensive Two-Dimensional Gas Chromatography Data Sets

Nadine Gawlitta(1), Uwe Käfer(1), Thomas Gröger(1), Ralf Zimmermann(1)*
(1) Joint Mass Spectrometry Centre of the University of Rostock and the Helmholtz Zentrum München, Munich and Rostock, Germany
**Corresponding Author: Nadine.gawlitta@helmholtz-muenchen.de*

Recently, two new commercially available software packages for the evaluation of two-dimensional comprehensive gas chromatography (GC×GC) data have been released. Both software packages allow peak alignment as well as advanced statistical and multivariate analysis within one software. Herein, a reference GC×GC data set, which has already been analysed via Statistical Compare (Leco ChromaTOF) and MatLab (MathWorks), will be re-analysed with the two new software packages. In this proof-of-concept approach, the results of the two new software packages will be evaluated and compared to the results obtained by the reference processing.

Introduction

Comprehensive two-dimensional gas chromatography (GC×GC) has technically matured over the last 15 years and became an important technique for the analysis of complex samples in industry and science. Possible fields of applications encompass environmental, petrochemistry and biological research and monitoring [1]. Depending on the processing approach (pixel or peak table based), an individual data file can consist of 10^2 to 10^8 different features and all of them can become variables in statistical evaluation [2, 3]. Moreover, for the comparison of different samples, peak alignment prior to data analysis is an indispensable step, as the



Analyte	^1t_R	$^1\sigma$	^2t_R	$^2\sigma$	C	F
27	1796	8,03638	0,7528	0,011644	25	0.19028
28	1805,69	10,4302	1,13306	0,014004	36	1.5947
29	1809,29	8,35711	0,651964	0,0106579	28	10.722
30	1815,75	9,90415	0,7195	0,0119097	20	1.7977
31	1816,83	9,32954	0,772	0,0145389	30	2.6016
32	1824,68	9,39286	1,26919	0,0404371	31	3.1199
33	1828,08	10,0077	1,21596	0,0185483	26	4.1918
34	1851,6	7,59934	0,8206	0,0109278	25	0.95551
35	1851,67	7,36788	0,873472	0,0139806	36	1.4961
36	1859,33	8,38033	1,2435	0,0190801	30	2.0900
37	1863,04	6,28627	0,842857	0,0108379	28	0.97603
38	1872,88	6,12759	0,810909	0,0107859	33	2.3328
39	1886,88	5,86302	0,824167	0,011948	24	3.1867
40	1884,06	7,00662	0,949844	0,00937433	32	6.6639
41	1897,74	10,4753	0,431129	0,0117409	31	6.0645
42	1895,42	5,77866	0,796667	0,0114018	36	2.1713
43	1916,71	3,19532	0,439857	0,0135302	35	9.4303
44	1903,55	6,08188	0,741774	0,0128159	31	0.85418
45	1907,6	10,5198	0,9212	0,15642	25	2.2713
46	1914,64	6,92935	0,594821	0,0126551	28	1.8695
47	1917,57	7,70518	0,711143	0,0113815	35	1.6490
48	1915,95	5,38958	0,787381	0,0147156	21	1.6966
49	1914,22	5,97162	0,895	0,0261509	32	0.076462
50	1916,94	5,51333	1,10014	0,0122174	36	1.1749
51	1923,46	7,17903	0,456346	0,0127717	26	8.6330
52	1926	6,39393	0,921571	0,0176056	35	1.2313
53	1930,43	3,96377	0,644348	0,0157613	23	1.4269
54	1927,71	4,90241	0,802571	0,0112702	35	1.7416

Figure 1: A) Depiction of the ChromaTOF build-in Statistical Compare. The sample set shows that the same analyte can have different retention times in first and second dimension, which explains the necessity of data alignment. ^1t_R (average retention time first dimension (s)), $^1\sigma$ (standard deviation first dimension (s)), ^2t_R (average retention time second dimension (s)), $^2\sigma$ (Standard deviation second dimension (s)), C (Count of samples, in which this feature was found), F (Fisher Ratio). B) Scores plot of PCA analysis proceeded via MatLab (MathWorks) and PLS-Toolbox (Eigenvector Research, Inc.).

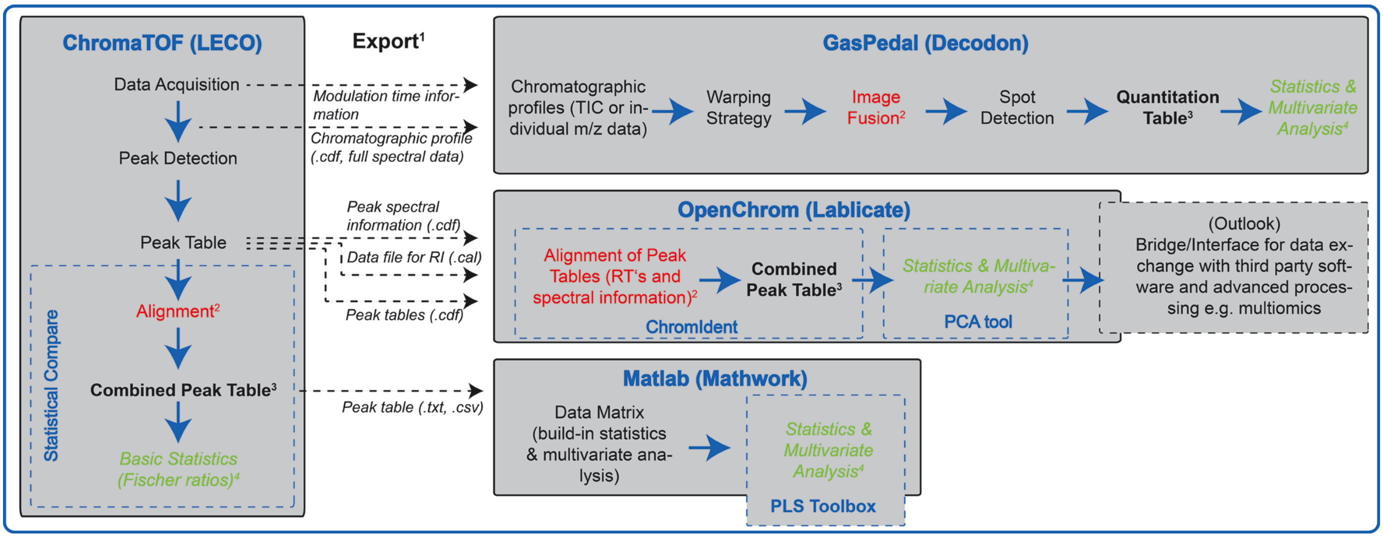

ChromaTOF (LECO)
Data Acquisition
Peak Detection
Peak Table
Alignment²
Combined Peak Table³
Basic Statistics (Fischer ratios)⁴
Statistical Compare
Export¹
Modulation time information
Chromatographic profile (.cdf, full spectral data)
Peak spectral information (.cdf)
Data file for RI (.cal)
Peak tables (.cdf)
Peak table (.txt, .csv)
GasPedal (Decodon)
Chromatographic profiles (TIC or individual m/z data)
Warping Strategy
Image Fusion²
Spot Detection
Quantitation Table³
Statistics & Multivariate Analysis⁴
OpenChrom (Lablicate)
Alignment of Peak Tables (RT's and spectral information)²
Combined Peak Table³
ChromIdent
Statistics & Multivariate Analysis⁴
PCA tool
(Outlook) Bridge/Interface for data exchange with third party software and advanced processing e.g. multiomics
Matlab (Mathwork)
Data Matrix (build-in statistics & multivariate analysis)
Statistics & Multivariate Analysis⁴
PLS Toolbox

*Figure 2: Leco ChromaTOF allows the **Export**[1] of data at different processing stages. A basic requirement for a comprehensive data analysis is a (semi-) automatic **Alignment**[2] of the individual data set to get a **Combined Data**[3] set for further comparative analysis. Significant features could be identified by **Statistics and Multivariate Analysis**[4]. TIC (Total Ion Chromatogram), RT (Retention Time).*

retention time of the same component can vary between samples. Herein, alignment of peak tables or chromatographic profiles are common. Subsequently, the generated data matrix is used for further statistical analysis to identify significant differences between samples or sample classes followed by multivariate data analysis. To date, no GC×GC supplier offers software that enables data alignment and comprehensive statistical evaluation. Recently, two new commercially available software packages have been released for the alignment and multivariate analysis of GC×GC data. In this article, these two software packages will be evaluated and compared to a reference data set analysed by Statistical Compare (Leco ChromaTOF) and MatLab (MathWorks).

Reference Data

The sample set chosen for statistical evaluation comprises a four week course of C3HeB/FeJ mice with a change in diet. Two different diets were applied, a high-fatty acid diet, referred to as safflower diet, and a low-fatty acid diet, referred to as standard diet. Therefore, liver samples of eight different classes (four weeks times two diets), with ten replicates each, were extracted and analysed with two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-ToF-MS) after derivatisation with N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA).

The generated data set was processed with Leco ChromaTOF (Version 4.32, Leco Inc) and MatLab (MathWorks). An alignment of peak tables was (semi-) automatically performed by ChromaTOF build-in Statistical Compare toolbox, which, additionally, offers the calculation of Fisher values (Figure 1A). For further statistical evaluation, the aligned peak tables were imported into PLS-Toolbox (Eigenvector Research, Inc), multivariate analysis clustered the samples into five different classes within the first two principal components (Figure 1B). Independent of the week of ingestion of the standard diet only one cluster can be observed (green), while the samples extracted from the mice that got the safflower diet can be differentiated in four groups depending on the time span of ingestion. A more detailed description of the study and further chemometric analysis of the data set can be found in the publication by Ly-Verdu *et al.* [4].

Methods

The original data set acquired with Leco ChromaTOF was re-analysed with two different software packages. GasPedal (Version 1.0.6) from Decodon is one of the two newly available software packages. Decodon, as a company, started about 20 years ago with the evaluation of two dimensional electropherograms (Delta2D) and has further developed since then [5]. Considering the similarity of two-dimensional profiles of GC×GC and two-dimensional gel electrophoresis, they started to dip into the field of two-dimensional gas chromatography. For data alignment, the chromatographic profile

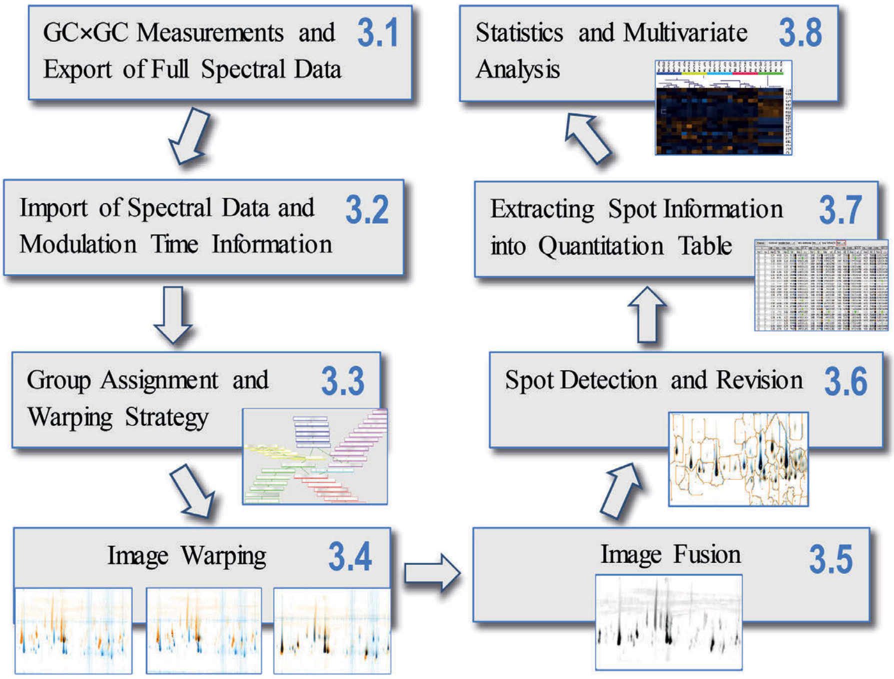

GC×GC Measurements and Export of Full Spectral Data 3.1
Import of Spectral Data and Modulation Time Information 3.2
Group Assignment and Warping Strategy 3.3
Image Warping 3.4
Image Fusion 3.5
Spot Detection and Revision 3.6
Extracting Spot Information into Quantitation Table 3.7
Statistics and Multivariate Analysis 3.8

Figure 3: Workflow of the software GasPedal.
Numbers 3.1-3.8 illustrate data import, alignment and analysis up to statistical evaluation.

as .cdf-Export (common data format; full spectral data) and the modulation time as complementary information has to be provided. The imported data sets are further processed similar to electropherograms and the processing follows the established workflows. Besides the chromatographic profile, GasPedal also uses the corresponding mass spectral information for the alignment, so called 'warping', of the two dimensional data (Figure 2). Schmarr *et al.* already investigated an early beta-version of GasPedal for the processing of GC×GC data generated from volatile fruit compounds back in 2010 [6]. In this article, a similar approach is used to introduce the workflow of GasPedal. After the export of the GC×GC data of ChromaTOF and the import of the required information into GasPedal, samples have to be assigned to groups and a warping strategy has to be set up (Figure 3.3).

The warping strategy depends only on the experimental setup, it can be chosen between Group, All-to-one, Chain and Group Chain Strategy. For the current data set, Group Chain Strategy was applied. Following the application of a suitable warping strategy, matching vectors between two GC×GC chromatograms are automatically found and defined by the 'Job Manager' when comparing two samples within each other.

An example of the warping of two different samples is demonstrated in Figure 3.4. The retention times of both samples (orange and blue spots) differ due to systematic deviation, thus, a direct comparison of the sample components is only liable after correct peak alignment (Figure 3.4). Warped images can be revised and manually adjusted, if necessary. Hereafter, all chromatographic images are fused to one artificial image. On this artificial image, peaks (also called 'spots') are detected and, to ensure that every important component is included for data evaluation, different fusion strategies can be applied (Figure 3.6). The results can be revised and spots can be added or deleted if erroneous spots have been detected. When the user is satisfied with the spots detected, the matching information is extracted and summarised in a quantitation table (Figure 3.7). This quantitation table builds up the base for further statistical analysis. For the identification of significant features, different filters can be applied and a t-test is automatically performed as last pre-processing step. Finally, various statistical applications like Analysis of Variance (ANOVA), PCA and hierarchical cluster analysis (HCA) can be used (Figure 3.8).

The second software that was investigated is OpenChrom from Lablicate. Lablicate has focused on the interpretation of one-dimensional gas chromatograms for many years and started to include the evaluation of GC×GC data about two years ago [7]. Their alignment strategy is similar to the one of Statistical Compare from ChromaTOF (Leco Peg4D). In the latest version available, individual peak spectral information (.cdf) and peak tables (.csv, comma-separated values) have to be imported (OpenChrom, Version 1.3.0, Figure 2). The release of the upcoming version will have replaced the import of .cdf-files by .csv-files only, including the mass spectral information within the same table. This change is expected to improve the performance of the software regarding import and alignment of data. For the export of data by Leco ChromaTOF, peak tables have to follow a distinct structure as listed below: Peak Number, Hit, Name, Classifications, Height, 1st Dimension Time (s), 2nd Dimension Time (s), Area and Spectra. An import of a retention indices data file (.cal, calendar scheduled data) is additionally possible. Data import and alignment are performed simultaneously resulting in a combined peak table. Matching restrictions (e.g. Minimum Matching Factor, Minimum Number of Ions, etc.) can be applied prior to the import. In the combined peak table, the occurrence of peaks in different samples can be compared. Mass spectral data of the peaks can be utilised to ensure correct data alignment. Prior to statistical evaluation, pre-processing in terms of excluding zero-values, normalisation (1-norm, 2-norm, inf-norm), transformation (Log 10 log(x), Power), centerng (Mean, Median) and scaling (Auto, Range, Pareto, Vast, Level) of the data can be applied. Different algorithms and filters are applicable and statistical implementations such as ANOVA and PCA can be used.

Results and Discussion

Due to time limitation, a small data set of the reference data consisting of twenty-five samples, which represents approximately 25% of the total reference data, has been selected and analysed via GasPedal. As an example of statistical evaluation, a hierarchical cluster analysis of this data set is shown in Figure 4. HCA classified the samples into five meaningful clusters, which correspond to the design of experiment. The first node already distinguished between the two diets, as the standard diet (green) showed less similarity to the mice samples with the safflower diet. The mice of the 'reverse' week got the safflower diet for three weeks and had one week of standard diet afterwards. The liver extracts of the 'reverse' week mice clustered with the samples from week 3. Week 1 and 2 built up another cluster. These results correspond to the ones published by Ly-Verdu *et al.*

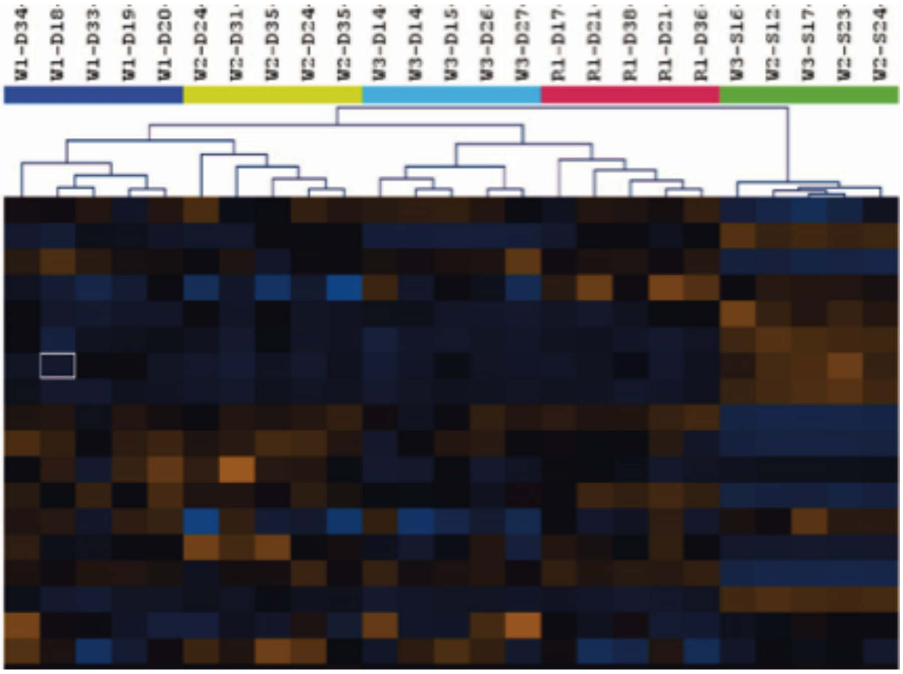


Figure 4: Hierarchical cluster analysis of a small data set of the reference data generated with GasPedal. Dark blue: Mice with high-fatty acid diet for 1 week. Yellow: Mice with high-fatty acid diet for 2 weeks. Light blue: Mice with high-fatty acid diet for 3 weeks. Red: Mice with high-fatty acid diet for 3 weeks and an additional week of low-fatty acid diet. Green: Mice with low-fatty acid diet.

[4]. Moreover, within this HCA, samples of week 1 seemed to be as similar to week 2 as samples from week 3 were to the 'reverse' week. A comparison of the cluster analysis published by Ly-Verdu *et al.* showed though that the samples from week 3 and the 'reverse' week grouped closer to each other than week 1 did to week 2. These deviations could be explained by the reduction of the data set to only twenty-five samples, which lead to less significant results.

For the evaluation of the second software package, the same selection of the reference data has been used. The statistical test focused on this time was ANOVA-PCA. Prior to the analysis via PCA, pre-processing in terms of normalisation, transformation, centering and scaling of the data was applied. Due to the pre-processing steps and the use of an ANOVA filter, supervised statistical evaluation could be accomplished. Four significant factors were extracted from the analysed data. Figure 5 shows the classification of the samples into five groups along the three most important factors. Factor 1, also referred to as Principal Component (PC), represents the difference of the samples regarding the time span of analysis. From right to left the weekly compositional change of the samples with week 1 (dark blue), week 2 (yellow) and week 3 (light blue) is illustrated. Additionally, the samples from the 'reverse' week (red) implied that the liver composition seemed to recover when food intake was changed back to standard diet, as they approach the sample composition of the mice that only ingested the standard diet (green). This observation equals the one made by Ly-Verdu *et al.* who also described a recovery of the liver samples of the mice that belonged to the group of the reverse week [4]. The PCA indicated that the time span of the intake of the standard diet has no influence on the composition of the liver extracts. These results deviate from the results presented by Ly-Verdu *et al.* [4]. This divergence could be explained by the reduction of the sample set to 25 samples only. Therefore, the individual variances of the samples had a stronger influence on the overall data analysis. For more significant data evaluation, the whole data set had to be studied.

Conclusion

Both software packages enabled advanced statistical and multivariate analysis of GC×GC-data. The produced results were comparable to the processing with Leco ChromaTOF build-in Statistical Compare and MatLab (MathWorks). OpenChrom as an established software provider for one-dimensional gas chromatography applied a similar alignment strategy as ChromaTOF and enabled comprehensive data analysis due to reasonable matching factors, detailed and replicable preprocessing opportunities and the applicability of various filters prior to multivariate analysis. GasPedal followed a different strategy and applied processing steps that are well established in routine analysis of electropherograms. Herein, images of two different samples were 'warped' in an acceptable way. Nevertheless, erroneous warpings and spot detection could be reviewed and corrected manually. Statistical and multivariate analysis strongly focused on the analysis of two-dimensional gels with many build-in statistical tools as hierarchical trees and PCA.

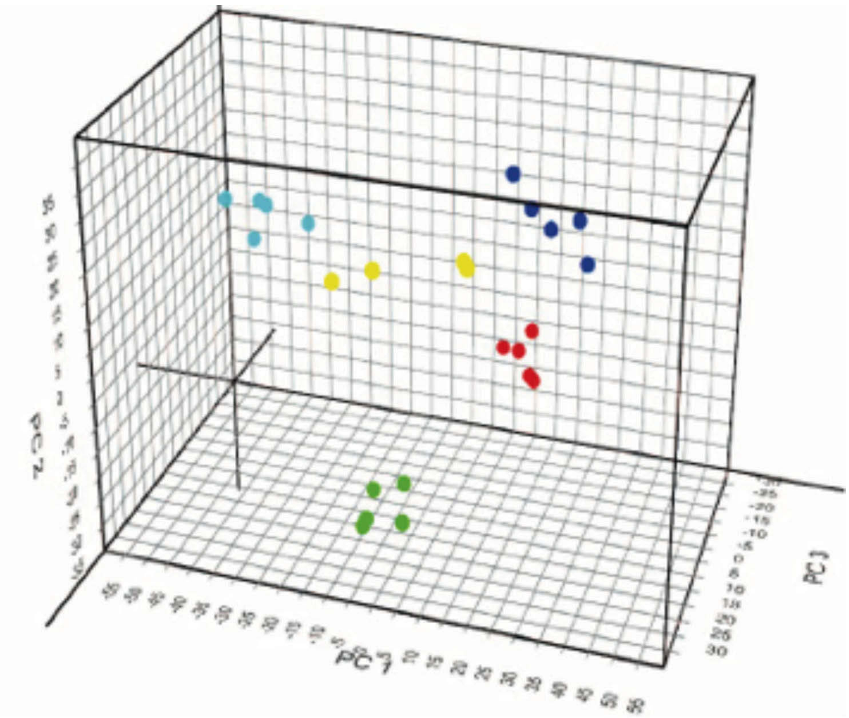


Figure 5: ANOVA-PCA of a small data set of the reference data analysed with OpenChrom. Dark blue: Mice with high-fatty acid diet for 1 week. Yellow: Mice with high-fatty acid diet for 2 weeks. Light blue: Mice with high-fatty acid diet for 3 weeks. Red: Mice with high-fatty acid diet for 3 weeks and an additional week of low-fatty acid diet. Green: Mice with low-fatty acid diet.

Acknowledgements

We gratefully acknowledge the Leco European Application and Technology Center Team (Berlin, Germany) for their ongoing support and cooperation. We would also like to thank Dr Andreas Klingberg (Lablicate GmbH, Hamburg, Germany) as well as Georg Wrettos (Decodon GmbH, Greifswald, Germany) for providing software and support.

References

1. Prebihalo, S.E., *et al.*, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications. Anal Chem, 2018. **90**(1): p. 505-532.

2. Pierce, K.M., *et al.*, Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts. Anal Chem, 2006. **78**(14): p. 5068-75.

3. Groger, T., *et al.*, Application of two-dimensional gas chromatography combined with pixel-based chemometric processing for the chemical profiling of illicit drug samples. J Chromatogr A, 2008. **1200**(1): p. 8-16.

4. Ly-Verdu, S., *et al.*, Combining metabolomic non-targeted GCxGC-ToF-MS analysis and chemometric ASCA-based study of variances to assess dietary influence on type 2 diabetes development in a mouse model. Anal Bioanal Chem, 2015. **407**(1): p. 343-54.

5. Berth, M., *et al.*, The state of the art in the analysis of two-dimensional gel electrophoresis images. Appl Microbiol Biotechnol, 2007. **76**(6): p. 1223-43.

6. Schmarr, H.G. and J. Bernhardt, Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques. J Chromatogr A, 2010. **1217**(4): p. 565-74.

7. [cited 2018 26.09.]; Available from: https://www.lablicate.com/#company.