TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Analytische Lebensmittelchemie

Deep Metabotyping of exhaled breath condensate (EBC) – characterization of surrogate markers for systemic metabolism and non-invasive diagnostics in Diabetes

Franco Moritz

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigte Dissertation.

Vorsitzender: Univ.-Prof. Dr. E. Grill
Prüfer der Dissertation:
1. apl. Prof. Dr. Ph. Schmitt-Kopplin
2. Univ.-Prof. Dr. M. Rychlik

Die Dissertation wurde am 18.09.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 28.03.2014 angenommen.

To my Family

And anyone who’s dear to me

Summary

Non-targeted metabolomics is a discipline of systems biology that has attained increasing interest in the recent decade. Metabolomics aims at the holistic and contemporary detection, quantification and identification of the entire set of small molecules that are being transformed or synthesized by living beings. The abundance of these metabolites varies over a range of several orders of magnitude and so do their physico-chemical properties. This strong variation makes it virtually impossible to fulfil the goal of metabolomics research in its entirety. It is either possible to detect, quantify and identify small sets of metabolites at a time (targeted metabolomics) or to maximize the set of detectable metabolites under loss of quantitative performance and possibility of identification (non-targeted metabolomics).

Non-targeted metabolomics however can extend the knowledge on biochemical processes, a feat that targeted metabolomics is barely capable of, as it can only revolve about existing knowledge.

We use non-targeted metabolomics under application of ultra-high resolution/accuracy mass spectrometry in order to determine the metabolic profiles of a yet poorly described sample matrix – exhaled breath condensate (EBC). EBC is an artefact of the pulmonary airway lining fluid and is therefore representative of its composition. While EBC research has focused on pulmonary diseases, we want to establish a link between EBC patterns and systemic metabolism. This endeavour is complicated by the exceptionally high variability in absolute metabolome concentration that is intrinsic to EBC.

The present thesis develops partially network based approaches for the extension of metabolite annotation and matrix effect control. Finally, it establishes the link between EBC and systemic metabolism at hand of the HuMet study.

Zusammenfassung

Ungerichtete Metabolomik ist eine Disziplin der Systembiologie, welche im vergangenen Jahrzehnt stark an Popularität gewonnen hat. Das Ziel der Metabolomik ist die holistische, zeitgleiche Detektion, Quantifizierung und Identifizierung aller kleinen Moleküle, welche von lebenden Organismen umgesetzt oder synthetisiert werden. Die Konzentration dieser Metabolite variiert über eine Spanne von mehreren Zehnerpotenzen. Gleichsam variieren ihre physiko-chemischen Eigenschaften. Diese starke variation macht es praktisch unmöglich die Ziele der Metabolomik – wie oben formuliert – zu verwirklichen. Es ist entweder möglich kleine Sets von Metaboliten zu detektieren, zu quantifizieren und zu identifizieren (gerichtete Metabolomik) oder es ist möglich – unter Verlust von quantitativer Performance und Identifizierbarkeit – das Spektrum detektierbarer Metabolite zu maximieren (ungerichtete Metabolomik)

Ungerichtete Metabolomik kann jedoch den Definitionsbereich biochemischen Wissens erweitern. Dies ist mit gerichteter Metabolomik kaum möglich, da sich diese nur innerhalb des bekannten Wissens bewegt.

In dieser Arbeit verwenden wir ultra hoch auflösende/akkurate Massenspektrometrie um mittels ungerichteter Metabolomik eine zur Zeit kaum beschriebene Analysematrix zu charakterisieren – Atemkondensat (EBC). EBC ist ein Artefakt der Oberflächenflüssigkeit, welche das pulmonale Epithelium überzieht. Während EBC Analyse bisher nur durchgeführt wurde um pulmonale Fragestellungen zu beleuchten, wollen wir in dieser Arbeit eine Verbindung zwischen EBC-Profilen und systemischem Metabolismus herstellen. Diese Anstrengung wird durch die starke variierende Konzentration des EBC-Metaboloms erschwert.

Die vorliegende Arbeit entwickelt Ansätze für die Erweiterung von Möglichkeiten der Annotierung sowie für die Kontrolle von Matrix-Effekten. Schließlich wird anhand der HuMet-Studie ein Link zwischen EBC und systemischem Metabolismus hergestellt.

Table of Contents

1 Introduction 11
1.1 Motivation and Overview 11
  1.2 Metabolomics 1: State of the Art and Theory 13
    1.2.1 Metabolism and ‘Omics 13
    1.2.2 State of the Art 17
    1.2.3 Chemo-Analytical Tools for Metabotyping 17
    1.2.4 Data Analysis 25
  1.3 Metabolomics 2: Practical Aspects – Closing in on Reality 26
    1.3.1 Practical Aspects of Instrumental Analysis 26
    1.3.2 Practical Aspects of Data Analysis: Understanding the Methods 27
  1.4 Exhaled Breath Analysis 40
  1.5 Diabetes mellitus 42
  1.6 Aim of the Thesis and Outline 44
    1.6.1 Aim of the Thesis 44
    1.6.2 Outlining the Thesis 44
2 Netcalc. A Network based Annotation Algorithm and its Adaptation to Metabolomics 47
  2.1 Prior to Annotation: Calibration and Error Distributions 48
  2.2 Annotation Strategies 51
  2.3 Netcalc 52
    2.3.1 The Concept 53
  2.4 Elemental Filters 54
    2.4.1 The Seven Golden Rules 55
    2.4.2 Senior’s Rules, the Cyclomatic Number and Extended Hybridization 57
  2.5 Metabolic REMDs 65
  2.6 Analysis of Error Sources 68
    2.6.1 Netcalc versus Database Matching 68
3 Data Analysis 1: Development of appropriate Workflows for Data Cleaning 77
  3.1 Data Handling 77
  3.2 The Matrix Effect 78
  3.3 Specifying the Problem 79
    3.3.1 Varying Feature Counts 79
    3.3.2 Power Relationships and Scale-free Relationships 80
  3.4 Analysis of Dilution Series 81
    3.4.1 Analysis of the six-point dilution series 81

3.4.2 Test of normalization strategy 95
3.5 Conclusion 97
4. Data Analysis 2: The Gauting Study 99
4.1 Analysis of data structure 99
4.2 Extraction of potential Surrogate Markers 102
4.2.1 Univariate analysis 102
4.2.2 Multivariate Markers 103
4.2.3 MassTrix annotation of Markers 105
4.3 Development of appropriate Workflows for the extraction and Interpretation of Surrogate Markers 107
4.3.1 The knowledge based strategy 110
4.3.2 The data-driven strategy 112
4.3.3 The intersectional approach 118
5. The HuMet Study: Links between EBC and Systemic Metabolism 122
5.1 Investigation of data structure 123
5.2 Elimination of binary bias 128
5.2.1 Multivariate Analysis 130
Comparison of eigenvectors with plasma Insulin, Glucose and Lactate 131
Grouping of Eigenvectors 134
5.3 Association to smoothed E-Blocks 136
5.4 MDEA of E-Blocks 139
6 Summary 141
6.1 Outlook 144
7 Supplementary Information 147
8 Bibliography 177
9 Acknowledgements 191
10 Scientific Communications 193
11 Curriculum vitae 195

List of figures

Figure 1: Scheme of the current disparity between research (left block) and customer (right block); colored according to existent or preferred dimensionality as indicated by the two ellipses 13

Figure 2: Schematic workflow of analysis in common metabolomics publications 29

Figure 3: Schematic representation of PCA with X being the data matrix, T being a set of so called t-scores, P^T being the transpose of a set of principal components (eigenvectors) and E being the residual between a prediction $\dot{X}$ and X, where $\dot{X}$ is TP^T. $\dot{X}$ does often not occur in such schemes (because some knowledge is implied here) 34

Figure 4: Second scheme supporting the understanding of PCA 34

Figure 5: Screen shots of the calibration process using a reference spectrum that was pre-annotated by Netcalc 50

Figure 6: Alginment of 15 mass spectra at the lower and upper end of the acquired m/z range. The m/z deviations are 0.07 ppm at m/z = 148 and 0.15 ppm at m/z = 655. Using the above introduced workflow the calibration of each spectrum took more or less 30 seconds. 51

Figure 7: Construction of mass difference networks. 53

Figure 8: Number of isobars per nominal mass under application of the relevant golden rules and under application of the adapted Senior rules. 64

Figure 9: Error-Mass plot Netcalc annotation of the Pubchem dataset. 65

Figure 10: Error over mass plot of MassTRIX annotation (0.1 ppm) of theoretical masses derived from previous Netcalc annotation. 69

Figure 11: Error over mass plot of MassTRIX annotation (3 ppm) of theoretical masses derived from previous Netcalc annotation. 70

Figure 12: Mass difference network. Two modules are visible. One module pertains to $[M+H^+]^+$ ions, the other module pertains to $[M+Na^+]^+$. Red edges highlighted by the blue ellipse are $[Na^+- H^+]$ REMDs, which allow the transition from proton space to sodium space. Black edges highlighted by the red circle are invalid connections between both ion spaces. 72

Figure 13: a) Counts of masses involved with false positive mass difference assignments (FP) and true positive mass difference (TP) assignments; a1) Greatest distance between TP and FP in metabolomics list; a2) Greatest distance of TP assignments between metabolomics and structural list; a3) TP to FP distance is increasing for the structural list as a function of networking error; a4) Increasing divergeance between FPs and TPs as a function of networking error in the metabolomics list. b) Metabolomics list| 0.5 ppm: Abundance of TP and FP mass difference assignments among reference mass difference groups [S = Sulphur Organic; P = Phosphorus Organic; O = Other Organics; KA = Keto Acids; ST = Structural; L = Fatty Acids and Isoprene units; DA = Dicarboxylic Acids; AA = Amino Acids]. c) Elemental compositions as a function of FPs and TPs. d) Red = Structural| Blue = Metabolomics| error distributions at 0.08 ppm networking error e) Red = Structural| Blue = Metabolomics| error distributions at 0.16 ppm networking error. f) Red = Structural| Blue = Metabolomics| error distributions at 0.5 ppm networking error.. 73

Figure 14: 10 different elemental combinations which almost exactly match the mass of [Na^+-H^+] – the transition from proton space to sodium space. The combination '-5C3H3O1P' is closer than 0.4 ppm to [Na^+-H^+] throughout the entire mass range. In addition, this mimicry moiety induces minimal changes in the isotopic pattern of higher masses because the isotope ^{13}C is not abundant enough to induce notable changes in the isotope pattern.. 74

Figure 15: 19 elemental combinations, which almost exactly match the zero mass. 75

Figure 16: Hypothetical Intensity-Frequency histogram referring to full ionization of 2000 metabolites (left) and Intensity-Frequency histogram referring to limited ionization yielding 1000 metabolites. The relative order of intensities in the samples is assumed to be constant. The red bar marks the median, green bars mark the first and the third quartile, empty bars relate to signals that get lost due to limited ionization. 79

Figure 17: Absolute Mass Defect over m/z plot. Red dots are 18 159 CHONSP compositions downloaded from Pubchem, corrected for proton removal; Blue dots are the EBC dilution series data. The orange Triangle indicates the forbidden zone................. 82

Figure 18: Running Pearson correlation coefficient over m/z. The yellow rectangle shows peaks with wrinkles or peaks splitting... 83

Figure 19: Intensity percentiles P = {50, 60, 70, 80, 90} over dilution factor D. The 100th percentile was excluded; the power distribution common to all mass spectral intensities made it impossible to plot the intensity profiles together.................... 84

Figure 20: Peak Counts over dilution factor D 84

Figure 21: Co-Intensity network of dilution set. Nodes are the m/z features and edges represent relations that satisfy cosφ >= 0.9. Nodes are colored according to modules (network sections which share more similarity among themselves than with the rest of the network). The network regions are characterized by three different themes: Theme A refers to 468 of 2178 (21.5%) intensity profiles that decrease with increasing dilution. Theme B refers to 53 of 2178 (2.4%) intensity profiles that strictly increase with increasing dilution. Theme C) refers to 1657 of 2178 (76.1%) intensity profiles that increase with increasing dilution but are barely present in 100% MeOH... 86

Figure 22: Absolute Mass Defect over m/z plot. The ellipsoid indicates the presence of salts. 88

Figure 23: Running Pearson Correlation Coefficient over m/z plot. 89

Figure 24: Running Pearson Correlation Coefficient over frequency of non-zero entries. 89

Figure 25: Characteristics of the Netcalc filtered dataset. A: The Error distribution of Netcalc annotations is well centered on zero ppm, which indicates precise and identical calibration throughout all 50 mass spectra in the set. B: In contrast to figure 22 the AMD over m/z plot is perfectly clean, which underlines the performance of Netcalc. C and D: The running Pearson Correlation Coefficient plots show, that Netcalc excluded co-linearity. 90

Figure 26: CIN of the HuMet-Dilution set. Nodes of the HuMet partition are red. The HuMet modules are well associated to the main dilution topic, which is theme C. 92

Figure 27: Different Modules of theme C that clustered in the co-intensity network and representative sum of scaled HuMet data. 93

Figure 28: HuMet modules normalized on representative data profile. 94

Figure 29: Zoom into the Ln($IQR_{CIN}/IQR_{Manhattan}$) percentile plot of the dilution dataset. Different cutoffs at 10%, 30%, 50% and 80% non-zero intensity frequency. Y-axis values < 0 indicate a smaller IQR_{CIN} than $IQR_{Manhattan}$ and the point of interception of the X-axis indicates the percentage of features for which Y ≤ 0 is true. 96

Figure 30: Eigenvector entries generated on the raw Gauting dataset plotted over frequency of m/z signals. The frequency of the largest positive and negative eigenvector entries is the maximum frequency, which indicates large intensities at large frequencies and the presence of different patterns in the data. 100

Figure 31: Eigenvector entries generated after normalization of rows, plotted over frequencies of m/z signals. The power distribution is cancelled out. The first eigenvector (E1), covering 43% of all eigenvalues, increases with increasing frequency. The other eigenvectors are independent of the frequency. 101

Figure 32: Sum of intensities over peak count. 102

Figure 33: PCA of the Gauting dataset differentiating smokers versus non-smokers over the eigenvector E205 and E608. 104

Figure 34: P-value distribution over the multivariately discriminating features 104

Figure 35: Log-Log plot of Rank of features sorted for decreasing connectivity (degree) over the degree itself. 100% relates to the full REMD set. 90% relates to 90% of the full REMD set (10% removed randomly) and so on. Linearity of the Log-Log plot indicates scale-freeness and a curved plot converges to Log-normality. A reduction of REMD number lets the degree distribution converge towards scale-freeness. 110

Figure 36: Log-Log plot of rank over degree derived from the Gauting network under use of the inner-net in comparison to the 5%, 10% and 100% of the full REMD set. The curvature of the inner-net is reduced in comparison to 10% and 100%. 112

Figure 37: Log-Log plot of REMD reduction by means of data driven MDEA. All three marker scenarios show a less random characteristic than the network based on the entire REMD set. The univariate data suggests less curvature and a higher maximum degree than the inner-net, which indicates a higher degree of organization. E205 shows a stronger curvature than inner-net but has a lower maximum degree, which indicates more random behavior. E608 has a more flat curvature but a lower maximal degree than the inner-net, which indicates more organization. 114

Figure 38: Vertical bar chart of z-scores for enrichment in univariate non-smoker (NS) and smoker (S) features and the respective REMDs. 115

Figure 39: Vertical bar chart of z-scores for enrichment in non-smoker (NS) and smoker (S) features along E205 and the respective REMDs. 116

Figure 40: Vertical bar chart of z-scores for enrichment in non-smoker (NS) and smoker (S) features along E608 and the respective REMDs. 117

Figure 41: Intersectional Approach. Two modules with most markers. Module 5 relates to non-smoker markers. Nodes pertain to highly oxidized molecules with most REMDs relating to glutamic acid derivatives, pimelic acids and adipic acid. Module 0 is composed of sulfur rich nodes and pertains largely to smoker markers. An indicator for an oxidative stress response module. 119

Figure 42: Balance between substrates and products in feature classes. 120

Figure 43: Plot of the first three eigenvector entries over the frequency of m/z values in the m/z-sample-intensity matrix. Large, positive eigenvector entries indicate a large degree of co-occurrence of an m/z with other m/z values throughout the samples. Large, negative entries indicate poor co-occurrence with other m/z values, i.e. singularities. 124

Figure 44: Sum of intensities over peak count. 125

Figure 45: A) Plot of the loadings of E1, E2 and E3 along the sample list. B) Plot of peak count and sum of raw intensities, both normalized on the maximum norm, along the sample list. 126

Figure 46: Eigenvector loadings of raw data. The large magnitude blocks relate to specimen 4, early specimen 6, specimen 7, specimen 10 and 14. 128

Figure 47: Normalization results over samples. 130

Figure 48: Three eigenvectors with the largest Pearson correlation towards the major clinical parameters determined by the HuMet consortium. 131

Figure 49: Eigenvector entries for the eigenvectors E4, E5, E6 and E10. Coloring of profiles according to HuMet challenge. 132

Figure 50: Eigenvector representative 133

Figure 51: Demonstration as to why smoothing enables the clustering of eigenvectors, which by definition are orthogonal. The linear equation y = 0.1*x was overlaid once with the trigonometric function y = 10*sin(0.1*x) and once with the function y = 10*cos(0.1*x). The pearson correlation coefficient of both functions naturally indicates orthogonality. Overlaying them with the linear equating gives a two functions whose oscillation center is monotonically increasing and that produce an r^2 of 0.18. We then performed average soothing. Each average encompassed 100 incident x variables. The result is a smoothed function with r^2 = 0.84. 135

Figure 52: Representative magnitudes of the E4|E6 block. There is a strong post-prandial response in SLD and OLTT, as well as positive entries in PAT (the positive entries at the begininning of OLTT are an artefact from smoothing). In addition, the late fasting phase seems to be co-directed with the other positve phases................. 136

Figure 53: Representative magnitudes of the E5 block. Strong positive responses occur almost exclusively in OGTT and OLTT. Markers related to this block can therefore be assumed to be markers of high carbohydrate and lipid loadings in blood plasma. 137

Figure 54: Representative magnitudes of the E8 block. Strong positive responses occur in the SLD and OGTT sections, exclusively. The fasting period, which covers 36 hours oscillates harmonically. 138

Figure 55: Data analytical workflow for deep metabotyping of EBC. The most important point is full control over, and removal of eventual biases. 143

List of Abbreviations

AA	Amino acid
AI	Aromaticity index
ALF	Airway lining fluid
AM	Adjacency matrix
AMD	Absolute mass defect
APCI	Atmospheric pressure chemical ionization
APPI	Atomsohperic pressure photo ionization
BUN	Blood urea nitrogen
CIN	Co-Intensity Network
CM	Co.Intensity Matrix
CoA	Coenzyme A
COPD	Chronic obstructive pulmonary disease
CSF	Cerebrospinal fluid
Da	Dalton
DBE	Double bond equivalent
DI-ICR-FT-MS	Direct Infusion ICR-FT-MS
DI-MS	Direct infusion mass spectrometry
DNA	Deoxyribonucleic acid
DRP	DNA→RNA→Protein
E1...En	Eigenvector 1 until eigenvector n
EBC	Exhaled breath condensate
EFE	Edge formation error
ELISA	Enzyme-linked immunosorbent assay
ESI	Electrospray ionization
FP	False positive
GC-MS	Gas chromatography – mass spectrometry
GDM	Gestational Diabetes mellitus
GSEA	Gene set enrichment analysis
HCA	Hierarchical cluster analysis
HMDB	Human metabolome database
HuMet	Human Metabolom Study
ICR-FT-MS	Ion cyclotron resonance fourier transform mass spectrometry
IE	Ionization efficiency
IQR	Inter quartal range
ISI	Insulin sensitivity index
Kegg	KyotoEncyclopedia of Genes and Genomes
KMD	Kendrick mass defect
LC-MS	Liquid chromatography mass spectrometry
LogD	Octanol/water partition constant at given pH
LogP	Octanol/water partition constant at the isoelectric point
LOWESS	Locally weighted scatterplot smoothing
m/z	mass over charge
MDEA	Mass difference enrichment analysis
MDN	Mass difference network
MeOH	Methanol
MS	Mass spectrometry
NADPH	Nicotinamide adenine dinucleotide phosphate
NMR	Nuclear magnetic resonance spectroscopy
NOM	Natural organic matter

OGTT	Oral glucose tolerance test
OLTT	Oral lipid tolerance test
OPLS-DA	orthogonal partial least squares discriminant analysis
PAT	Physical activity test
PCA	Principal component analysis
PLS	Partial least squares or projection to latent structures
pH	Activity of hydronium ions in aqeous solution
PP	Pyrophosphate
ppm	Parts per million
Q-MS	Quadrupole mass spectrometer
REMD	Reaction equivalen mass difference
R	Resolution
RNA	Ribonucleic acid
SLD	Standard liquid diet
S/N	Signal to noise ratio
SOM	Self organizing maps
SPE	Solid phase extraction
TCA	Tricarboxylic acid
TOF-MS	Time of flight mass spectrometry
TP	True positive
V	Volt
VOC	Volatile organic compound

1 Introduction

1.1 Motivation and Overview

The title of the present manuscript contains four major keywords, which are ‘Metabotyping‘, ‘Surrogate Marker‘, ’Exhaled Breath Condensate‘ and ‘Systemic Metabolism’. Metabotyping is a sub-discipline of ‘Metabolomics’ and/or ‘Metabonomics’ (both increasingly used synonymously) [Nicholson, J. K., et al., 2008]. Both techniques are based on the parallel measurement (quantitation and/or identification) of large sets of molecules, and therefore employ multivariate statistics for data management [Lindon, J. C., et al., 2008]. The term ‘Systemic Metabolism’ [Soininen, P., et al., 2009] refers to metabolic processes, which pertain to an entire organism in contrast to the metabolome of histologically defined tissues. ‘Surrogate Markers’ [Kumar, M., et al., 2009] are defined to be (any) analytical measures, which allow statements on a process or state of interest without having to invasively dissect (and therefore disturb) that process or state itself.

Surrogate markers can be the absolute levels of analytes or relative concentrations or patterns of analytes, which can significantly and validly be associated to e.g. a disease, nutritional state, state of health of an eco system and many more. Ultimately, it is as well desirable for a marker to enrich current knowledge on a metabolic state and for it to point out means of treatment. An in-depth introduction on surrogate markers and metabotyping is given in chapter 1.1.

A prerequisite for the definition of surrogate markers is the definition of a chemo-analytical workflow.

Workflows in traditional physiological research are deductive, i.e. hypothesis driven. The use of deductive workflows is appropriate if there is a fundamental body of knowledge, which supports the hypothesis that the workflow is supposed to test or verify. For example, if there are numerous indications that a specific biochemical pathway (like cholesterol biosynthesis) is associated to heart disease, it is reasonable to choose a set of analytes related to the respective pathway and to set up an analytical strategy, which is optimal for the sensitive detection and quantification of these targets.

However, if only the symptoms of a disease are known, if indications from literature do not converge, or if a disease may be too multi-factorial to comprehend, it can be more useful to choose an inductive route. Inductive science in the context of metabolomics and metabotyping is often referred to as being ‘non-targeted’ for the fact that no analytical target is defined;

notwithstanding that there actually is a hypothesis underlying inductive research.

This hypothesis could be formulated as follows:

'Given the existence of at least two different populations of metabolic phenotypes among a sample set, there should be at least one feature among all measured variables, which discriminates between the metabolic phenotypes. Likewise, there might be a set of such features, which enables the distinct determination of a metabolic phenotype'

This hypothesis implies the existence of a chemo-analytical technique that is in fact capable of not only the detection, but as well the resolution of a multitude of such features.

In the present thesis, surrogate markers for systemic metabolism – and optimally diabetes mellitus – are to be found in exhaled breath condensate (EBC), which is a surrogate matrix for the airway lining fluid (ALF) [Hunt, J., 2002]. The ALF guarantees the optimal mechanical function of the lung and supports the molecular intercourse between an (aerobic) organism and its environment.

In the search for surrogate markers, EBC has almost exclusively been investigated in the context of pulmonology and clinical chemistry. Literature on EBC analysis reflects the impact of clinical chemistry, as the entire spectrum of so far analyzed compounds is a subset of standard clinical determinants for inflammatory actions and – more specifically – pulmonary complications [Risby, T. H., et al., 1999; Cao, W., et al., 2006]. The concern that analyte patterns in EBC could reflect diseases in other organs or the nutrition state of a human being were never seriously formulated even though exhaled volatile organic compounds (VOCs) such as acetone are known to reflect the metabolic state of the liver, the kidney or the intestine [Phillips, M., 1992].

Current metabolomics preferably uses multi-variate techniques of data analysis rather than uni-variate data analysis. Multi-variate data analysis is performed by analyzing measures of similarity or distance between the relative (also semi-quantitative) abundances of all variables in a given dataset. In contrast, uni-variate data analysis concerns with the comparison of statistical moments such as mean, standard deviation, variance or median and inter quartile range between two populations of samples (e.g. healthy versus diseased). Using such techniques, an analyst wants to infer about the significance of the difference of absolute variable concentrations.

Studying the surrogate marker catalogues of clinical medicine laboratories (e.g. of the Charité in Berlin) shows that applied markers are exclusively uni-variate in nature, which reflects the fact, that no multivariate surrogate marker has made it to clinical practice within two decades of ‘omics science.

Figure 1 gives a schematic representation of the disparity between science on surrogate markers and the end-users requirements (pharmacy, clinical chemistry, the patient) of markers. Here may lay one of the major causes for which ‘omics sciences have yet to deliver their first accepted clinical marker.

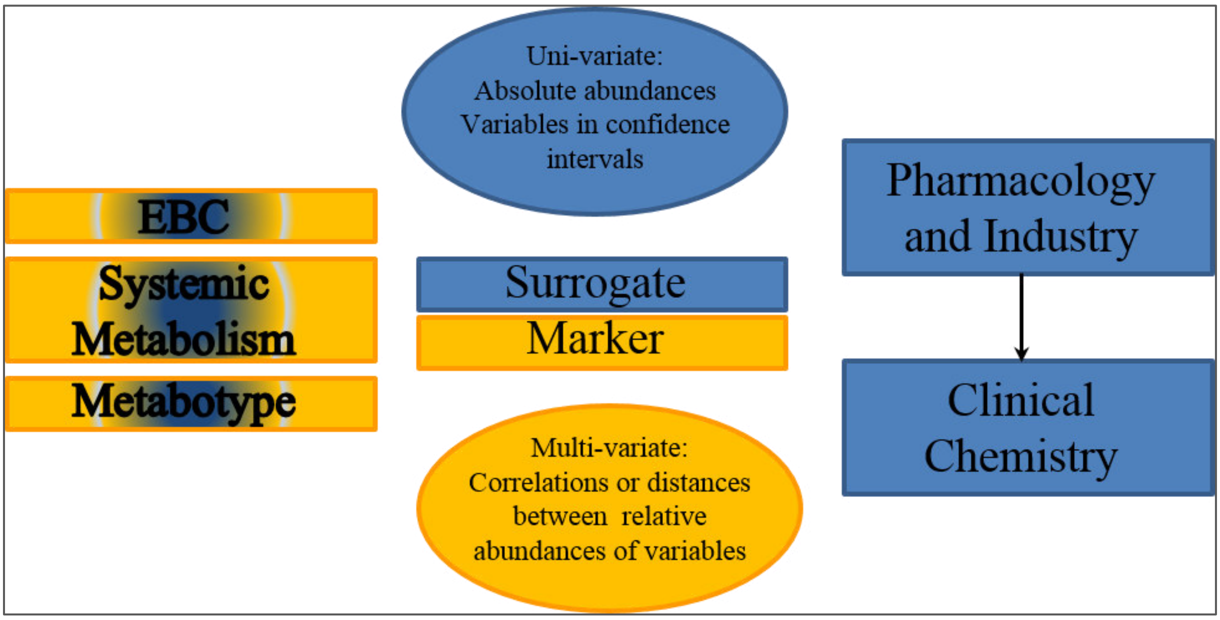

Uni-variate:
Absolute abundances
Variables in confidence intervals
EBC
Systemic Metabolism
Metabotype
Surrogate
Marker
Pharmacology and Industry
Clinical Chemistry
Multi-variate:
Correlations or distances between relative abundances of variables

Figure 1: Scheme of the current disparity between research (left block) and customer (right block); colored according to existent or preferred dimensionality as indicated by the two ellipses

An introduction into different techniques of data analysis will be given in chapter 1.3. The reader is introduced into characteristics of EBC data in chapter 1.4. Chapter 1.5 summarizes the previous sub-section and points out the major problems which have to be treated in this thesis.

1.2 Metabolomics 1: State of the Art and Theory

1.2.1 Metabolism and ‘Omics

Metabolism is a term that pertains to change; the (inter)conversion of organic substances (metabolites) is a function and condition for what is called “life”. Metabolism contrasts living beings from non-living, abiotic things. The question as to what “life” is has puzzled mankind since its existence. Up until the end of the 19th century, “life” was something vaguely described, something that fell under the realm of vitalism, i.e. processes that characterize “life” have an inherent and mystic “life-force”. Finally, in the last century, fast progresses in biochemistry and/or physiology, (cell)biology and many more have yielded a more objective definition as to what “life” actually is: According to [McKay, C. P., 2004; Davison, P. G., 2008] life is characterized by objects that

- are composed of one or more cells, which
    - maintain homeostastis; a constant inner state (like the relatively constant organo-chemical setup of a cell)
    - grow and/or reproduce themselves (autocatalism)
    - have the ability to adapt their homeostatic state to environmental changes
- respond to stimuli
- “perform metabolism”

So metabolism is the set of chemical conversions, which fuels or enables all the other manifestations of life. It encompasses

Catabolism:	the decomposition of organic compounds into smaller organic compounds. This process produces and fixes energy and small organic molecules that can be used in anabolism.
Anabolism:	the process that uses the energy and small molecules produced in catabolism in order to build larger molecules.

These larger molecules are in turn DNA, RNA, proteins (polypeptides), carbohydrates (polysaccharides) and lipids of different kinds. DNAs, RNAs and proteins – enzymes specifically – curiously contain the “blue print” for metabolism in that DNAs code for RNAs and RNAs code for proteins (enzymes inclusively) and then the coding stops. Proteins, carbohydrates and lipids in turn build up physical structure. The compartmentalization of a cell – much like a funnel – directs and optimizes the flow of mass and energy.

So the living cell is the manifestation of its own intertwined and inter-causative actions. Omics sciences – genomics, proteomics and metabolomics in particular – make use of instrumental-analytical techniques and ever more powerful computers and computer science in order to study the concerted responses of self-interactive living systems towards stimuli. Therefore, these scientific disciplines are summarized under the term “systems biology” [Villas-Boas, S.G., et al., 2007; Nicholson, J., 2006]

.

As indicated above, genomics and proteomics are “coding” each other and can therefore be directly compared. That means the amount of RNA transcripts stemming from the DNA template can be compared to the amount of protein that it codes for and the coded pairs can directly be associated to each other. In that sense, cause and response can unequivocally be mapped to each other. These properties of DNA, RNA and proteins are reflected in their instrumental analysis; DNA and RNA can be amplified in abundance and sequences and the protein's abundance can be measured (e.g. by means of mass spectrometry) and their code can be sequenced.

The subjects of metabolomics do not have such direct coding towards DNA, RNA and proteins; rather their template is implied or manifests in DNA, RNA and proteins, by the specific functions that proteins have on metabolites. This again is impressive: Physical cell structure, enzymes and DNA codes; everything is constructed around the virtual image of metabolic reactions that again serve for the self-maintenance of the very same construction.

The "mapping domain" is the domain of genotype, the non-mapping domain (the metabolites) manifests the phenotype.

However, this indirect implication of a metabolite in DNA, RNA and proteins complicates the analysis of metabolites; their structure cannot be "physically mapped" against the proteome and the genome; they can only be mapped by virtue of function, i.e. the observation of changes in phenotype as a function of changes in genotype. Instrumental analysis of metabolites is therefore inherently dependent on an experimental setup in which one group of cells is allowed to stay "normal" and another group of cells is perturbed by a specific stimulus. This stimulus can be a change in nutrient composition or a genetic manipulation. This experimental approach is the deductive (targeted or hypothesis driven) approach. The inductive approach (non-targeted or data-driven) would be the collection of individuals of different phenotype and the consequent differentiation of their genotype. Interestingly, the inductive approach is poorly accepted among scientists, even though the achievements of Gregor Mendel were based on inductive experiments and not on deductive experiments.

Metabolomics is ultimately the static or dynamic description of a living system's molecular phenotype (metabotype) [Nicholson, J. K., et al., 2002; Holmes, E., et al., 2008].

Metabolic Pathways

Throughout the taxonomic system of biology, each genotype defines a species but different phenotypes develop as a function of environmental stimuli – weather, nutrition, parasitic interactions between species (a form of disease) [Gavaghan, C. L., et al., 2000]. Phenotypes of a species can also change as a function of slight variation in genotype, which can be normal and it can also be a form of disease. As indicated above, metabolism is multifactorial, which complicates its analysis. Yet, metabolism adheres to the organization of cellular structure and the enzymes it contains. Metabolism itself is therefore organized and its organizational subunits are called metabolic pathways. Metabolic pathways describe the sequence of reactions, which a living cell performs in order to convert a metabolite A into a metabolite Z or into a poly-Z structure. Commonly, on the way from A to Z energy is either produced and stored (catabolism) or energy is required to build up structures (anabolism). The first pathways discovered were glycolysis, the Calvin cycle and the tri-carboxylic acid cycle.

Metabolic pathways are entirely anthropogenic; they indicate directions and connections between entities that were significantly associated in the human eye. It is not known whether each of the known pathways is naturally “intended” as it is described by scientists. However, in the end pathways are the major observed routes of mass flux that are addressed by a (experimental) stimulus.

The analysis of metabolic pathways supports the classification of genotypes and phenotypes because they represent an ordered system or network whose topology can be more or less specific for a species. While the primary energy metabolism throughout different species is very similar, the secondary metabolism – that what happens with metabolites apart from energy production – may vary strongly, and may thus finally be the determinative difference between phenotypes.

Also, the magnitude of mass flux that an organism directs through a pathway is indicative of the “metabolic preferences” an organism, tissue or cell has. As a consequence, even if pathways are hypothetical (often experimentally verified) constructs, which enable a scientist to systematically compare different species and phenotypes; they – much like a road map – help scientists to communicate the site and direction of an event.

The sequence of enzymes along a metabolic pathway can be used to relate the metabolome to the proteome and the genome [Nagarajan, N., et al., 2010; De Souza, A. G., et al. 2009].

Uses of Metabolomics and Omics in general

Apart from gaining an understanding of the things, these disciplines’ major purpose is the identification of control points, or surrogate markers, which can be used to judge, whether a process works as it should and if not, why it does not. This knowledge again should enable the human to control the process and revert it into a normal working state.

Consequently, metabolomics and/or metabotyping is of large interest for medicine, biology and ecology. Human action has caused several “abnormal” developments in ecosystems, in the environment in general, and the increasing amount of humans on earth nurtures the

development of new diseases. Humans have an interest to control all these factors.
As indicated above, what is necessary for control is an understanding of what is to be controlled in order to manipulate a process most effectively. The formulation of surrogate markers is one central part of this thesis.

1.2.2 State of the Art

Omics sciences gained much attention throughout the last decade and metabolome analysis is widely applied. As it was recognized, that symbiotic microorganisms in the human intestine have the ability to modulate human metabolism, there is large interest in the deconvolution of the inherent regulatory mechanisms [Nicholson J.K.et al. 2005]. Metabolites, which are produced not by the host himself were decided to be called co-metabolites [Li, M., et al., 2008]. It is generally acknowledged, that the microbial setup of intestinal flora (the Microbiome) is able to modulate the physiological state of a host; e.g. Crohn’s Disease [Jansson et al., 2009]. A broad multi-platform screening of human nutritional metabolism was recently published in the scope of the HuMet study [Krug, S., et al., 2012]. Here 15 volunteers were led through multiple nutritional challenges and their blood plasma, urine and EBC were analyzed by means of enzymatic assays (ELISAs), NMR, LC-MS and *Ion Cyclotron Resonance Fourier Transform Mass Spectrometry* (ICR-FT-MS). In close relation to nutritional habits and intestinal microflora, diabetes research is in the focus of metabolomics endeavors.

1.2.3 Chemo-Analytical Tools for Metabotyping

Metabolomics, in particular was on its way since the 1980’s, where NMR experiments on complex mixtures were extended to a broad scan-biochemistry concept by the Nicholson laboratory [Nicholson J.K et al., 1983; Nicholson, J.K et al., 1985; Bales J.R.et al., 1984; Gartland K.P.R., et al., 1989; Nicholson J.K., et al., 1989, 1989; Moka D.et al., 1998]. Finally, in 1999, the term ‘Metabonomics’ was born [Nicholson, J.K et al., 1999]; Olivier and Fiehn defined ‘Metabolomics’ in 2000 and 2001, respectively [Fiehn, O., 2002]. Both disciplines have inherently the same aim: a broad band detection and description of the response of as many as possible (if not ALL) metabolites in a living system towards stimuli. Nicholson had pharmacological stimuli and diseases in mind, Olivier and Fiehn focused on plant manipulation. The research unit Analytical BioGeoChemistry at Helmholtz Zentrum

München – led by Prof. Philippe Schmitt-Kopplin – has a vast body of experience in the field of complex mixture analysis, which roots in the analysis of natural organic matter. In this discipline, high resolution techniques, which are drafted in the following sections, build a fundament for the detailed description of compositional- and chemical spaces of any sample type [Hertkorn, N., et al., 2008]. Compositional metabotypes, as they are analyzed in the present manuscript, envelope the structural spaces of the known metabolome, and they therefore include the unknown metabolome as well.

1.2.3.1 Physical Principles of Measurement

NMR

The nuclei of elemental isotopes of odd neutron number have a non-zero electro-magnetic spin, which causes these nuclei to oscillate at a specific frequency in a homogenous magnetic field of a given strength. Depending on the immediate stereochemical environment of such an isotopic atom, oscillation frequencies deviate from normal frequencies of whichever reference compound. In order to acquire NMR spectra, all analytes need to be able to oscillate freely, which requires liquid samples. By means of magic angle techniques, intact tissue samples can be analyzed. However, tissues can only be obtained by means of invasive biopsies.

Since NMR depends on the existence of rare nuclei with odd neutron numbers, this technique is inherently insensitive. Consequently, it is only possible to acquire spectra pertaining to the most abundant metabolites or to acquire spectra of purified substances. A technique, which is less quantitative and specific but more sensitive is mass spectrometry.

Mass Spectrometry

All elements and each of their isotopes have a well defined mass. Isotopes of an element have equal numbers of positrons and electrons but different numbers of neutrons. The most abundant isotope of an element is taken to be the reference isotope which represents that element. The percentage of less abundant isotopes may vary depending on age and origin of the element (e.g. relative abundances are different in meteorites or meteorite craters than they are in the earth's crust). Also depending on the decay of an isotope like ^{14}C, sequestered moieties like deep see organic matter may have different abundances of this isotope as compared to the earth's crust [Flerus, R., et al., 2012]. These exceptions put aside, relative abundances can be seen as fairly constant.

A molecule only composed of the most abundant reference isotopes is said to have the *exact* *mass* as opposed to the *molecular mass*, which is the weighted average of all stable isotope permutations of all the elements which make up a molecule's *elemental formula* or *sum formula* (e.g. Glucose: $C_6H_{12}O_6$). A molecule can as well be attributed a composition, which is the smallest divisor of a formula (e.g. Glucose: $C_6H_{12}O_6 = 6*C_1H_2O_1$). Back in the days of Justus von Liebig, chemicals were described by pyrolysing a sample and measuring the amount of e.g. C, H and O. In these days large amount of pure sample were needed, so that their remains could be weighted by means of a common laboratory scale. By modern standards, this method is very insensitive and inaccurate since much more accurate measurements can now be carried out by means of mass spectrometry.

Mass spectrometry being a much more sensitive and accurate technique than mechanical scaling of a sample's weight, can use the above described relative isotope abundances as a means of molecular formula validation.

The first concept in mass spectrometry is that a charge that is transferred to a molecule – using one of a variety of techniques to be introduced later on – attributes the molecule with a mass to charge ratio (*m/z*). Once exposed to an electric or magnetic field in an evacuated chamber, it is possible to accelerate or decelerate and to manipulate the trajectory of a charged molecule. Because of the principle of mass inertia and the quantized nature of charges, equal masses of the same charge state are experiencing the same force when being exposed to an electric or magnetic field of the same strength. Mass inertia then causes molecules to have different final linear velocities and different electromagnetic deflections. All known measures for the differentiation of m/z ratios are proportional to time, which will later on turn out to be an important note.

Depending on the physical concept and the architecture of a mass spectrometer, the response to the manipulation exerted on molecules of similar yet different *m/z*, have different magnitudes. The magnitude of response also varies as a function of the ion number populating the mass spectrometer at a time. After manipulation of a sample of molecular ions their response is commonly measured by recording an image current, which is generally produced by letting the ion flows pass by a transistor.

The detected responses to manipulation are finally transformed into a spectrum by applying the mathematical relation which describes the ion's behavior in the mass spectrometer and by plotting the resulting m/z and magnitude of response on the spectrum's x-axis and y-axis, respectively. The strength of response is typically proportional to the abundance of ions of the same m/z in the mass spectrometer; it does not necessarily imply proportionality to the analyte's abundance in a sample.

The quality of a so produced mass spectrum can be assessed by the following measures:

Mass Accuracy: Mass accuracy is defined to be the absolute or relative difference between the theoretical mass of a compound with a given sum formula and the m/z ratio measured (normalized to z = 1).

Absolute Mass Accuracy = $(m_{measured}-m_{theoretical})$ in Da.

Relative Mass Accuracy $=10^6*(m_{measured}-m_{theoretical})/\ m_{theoretical}$ in parts per million (ppm).

Mass Resolving Power: Mass resolving power reflects the conciseness of the separation of two adjacent *m/z* species, which in default of an adjacent *m/z* species is expressed as the given *m/z* value standardized on its full width at half maximum peak height:

Resolving Power (R) = $[m/z]/\Delta[m/z_{50\%}]$.

Sensitivity: The sensitivity of a mass spectrometer is always linked to the capability of the instrument to produce an m/z signal of one compound larger than a specified signal to noise ratio. Sensitivity is therefore compound specific and is not an inherent measure of a mass spectrometer's quality. It much rather reflects the efficacy of an analytical procedure from sampling through sample pre-treatment up until mass spectrometric measurement.

Signal to Noise Ratio: A mass spectrometer interacts not only with ions but with any kind of electromagnetic irradiation as well. These and other interactions cause a base-line response of the mass spectrometer, which is called noise. It can be attributed with a standard deviation if it is Gaussian. The signal to noise ratio (S/N) expresses the distance of an m/z peak magnitude from the mean noise level in quants of standard deviations. An S/N major to 2.5 indicates a deviation from the noise distribution with a probability $P < 1\%$ for the peak to be a random aberration from the noise level.

Duty cycle: The duty cycle is the time that a molecule takes to cross all stations of measurement: ionization → ion optics → "mass separation" → scan → computerized transformation of the signal into a mass spectrum. It is an important measure for the adjustment of mass spectrometric measurement counts (sampling rate) to the resolution of hyphenated chromatographic techniques.

Ionization of the sample matrix

For mass spectrometric measurements, it is necessary to transfer analytes from liquid phase into gas phase and to ionize them in that process. Nowadays, the principal concept for ionization is based on pumping the liquid sample through a thin metal tube and spraying it into a heated ionization chamber. Ionization is then realized by different principles. In atmospheric pressure photo ionization (APPI) the liquid sample partition is completely vaporized at high temperatures and ionization is induced by the impact of ultra-violet light. This process leads to electron abstractions in the π-systems of analytes. Atmospheric pressure chemical ionization (APCI) is based on complete liquid vaporization as well. However, ionization works by arcing or corona discharge on the tip of a metal needle that is placed in the ionization chamber. This discharge ultimately transfers electrons onto the gaseous environment in the ionization chamber, which results in radical ions. These radicals then ionize the analytes.

The third technique, which is softer than APPI and APCI, is electrospray ionization (ESI). It is the most frequently used ionization technique in the mass spectrometry of liquid samples. A comparison of all three techniques in terms of sensitivity and analyte specificity is provided in a pharmaceutical study [Garcia-Ac., A., et al., 2011]. It was found that ESI ionization is more sensitive towards phosphocholines and sphingomyelins while APCI was more sensitive towards phosphoethanolamines [Byrdwell, W. C., 1998].

Electro Spray Ionization (ESI)

ESI is the most used method in metabolomics, as it coveres a wide range of analyte specificities, and since there is fewer analyte fragmentation than in other techniques. In ESI, ions are produced in solution while the sample is sprayed through a grounded metal capillary, which is placed in vicinity to the charged mass spectrometer entrance. This setup generates an electric field and ions are separated in the tip of the ESI capillary. As a consequence, liquid surfaces are populated by charges of the same polarity. Their repulsion causes explosions of liquid droplets in the sprayed sample while a heated gas stream vaporizes the solvent. The vaporization causes the charged sample droplets to shrink, which in turn increases surface tension and charge density. Eventually the charged droplets will explode again due to coulombic repulsions.

Common electric field strengths which are applied for ionization vary between 3,000V and

4,500V, which causes an electric field of 1,000V/cm [Chech, N. B., et al., 2001].

The ESI process can produce ions by means of proton abstraction or clustering with anions in negative mode ($[M-H]^-$, $[M+F]^-$, $[M+Cl]^-$) and by the formation of different clusters in positive mode ($[M+H]^+$, $[M+Na]^+$ or $[M+K]^+$) [Boutegrabet, L., et al., 2012]. Positive ionization is thermodynamically favored as less energy is needed for adduct formation than for breaking a covalent bond. Different compound classes have different ionization efficiencies (IE) given the same conditions because they differ in polarizability, their distribution coefficient between an aqueous solvent and hexane ($LogP_{Hexanol}$) and they differ by their pK_a [Oss, M., et al., 2010; Cole, R. B., et al. 1993; Henriksen, T., et al., 2005].

It has been shown that the affinity of N-hetero aromatic compounds to $[M+H]^+$ ion production is five-fold larger than the affinity of oxidized polyaromatics towards $[M+H]^+$ ion production [Oss, M., et al., 2010]. It was as well shown, that LogP has a larger influence on ionization than acidity. In consequence, surfactant molecules tend to suppress other signals both, in negative and in positive ionization mode [Cole, R. B., et al. 1993; Henriksen, T., et al., 2005]. On the other hand, a reduction in ES droplet size can compensate these effects. Respecitve nano-ESI sources are available, but they are difficult to handle because their small dimensions support clotting and small deviations in the used material can cause stronger changes in responses as compared to conventional ESI.

In consequence, if the aim of a metabolomics study is to maximize metabolically relevant information, the spiking of standards into samples (for calibration purposes) has to be avoided when direct infusion mass spectrometry is applied.

Mass Spectrometers commonly used in Metabolomics

The most common mass spectrometers in general are variations of the quadrupole mass spectrometer (Q-MS) and the time of flight mass spectrometer (TOF-MS). A less common but unequally stronger mass spectrometer in terms of accuracy and resolution is the *Ion Cyclotron Resonance Fourier Transform Mass Spectrometer* (ICR-FT-MS). Since 2006 another Fourier transform mass spectrometer has entered the market, the Orbitrap. In terms of resolution and accuracy it is placed between Q-MS/TOF-MS and ICR-FT-MS.

Predominantly used for metabolomics experiments – in conjunction to chromatographic techniques – is the TOF mass spectrometer. Orbitrap and ICR-FT-MS are less commonly used; the first because of its young existence, the second because of its requirements in terms of laboratory space and its expensive price. The next pages will give a short introduction into these mass spectrometers' concepts. The aptitude of these apertures for metabolomics experiments is afterwards discussed in Metabolomics 2.

Time of Flight Mass Spectrometry (TOF-MS)

TOF mass spectrometers measure the flight time that an ion needs in order to pass a field-free zone, which is called 'the flight tube' [Guihaus, M., 1995; Mamyrin, B.A., 2001]. Prior to the flight in the flight tube, charged molecules are accelerated by an electric field. As all ions of the same charge receive the same force, ions of different mass reach different terminal velocities.

The following equation describes how mass (m), charge (z), field strength (eV), and flight tube length relate to each other in a TOF mass spectrometer.

$$tof = \frac{L}{v} = L\left(\frac{m}{2zeV}\right)^{\frac{1}{2}}$$

The time of flight increases linearily with the length of the flight path, which causes a similar increase in resolution. In a review on mass spectrometric techniques we have calculated that a Bruker MaXis3G-TOF can perform 10.000 consecutive scanning events per second, given m/z = 1000 [Forcisi, S., et al., 2012]. A reviewer of the manuscript had pointed out, that the length of duty cycles of TOF mass spectrometers is determined by the accumulation of scans, which increase sensitivity.

The aptitude of a mass spectrometer for metabolomics measurements is determined by their resolving power, accuracy and precision. Most efforts in the development of TOF mass spectrometers were centered on the increase of flight paths in the last decade. For this reason, different techniques for flight path reflection were developed. An extreme example is the high resolution TOF developed by LECO Corporation, which provides a resolving power of 100.000 at m/z 400 [Klitzke, C. F., et al., 2012]. TOF mass spectrometers are sensitive to temperature insulation and all mass spectrometers can potentially be over-saturated by too high ion abundances. However, TOF mass spectrometers are commonly more resistant to oversaturation than the high resolution ICR-FT mass spectrometer, which was used in this thesis.

Ion Cyclotron Resonance Fourier Transform Mass spectrometry (ICR-FT-MS)

Ion Cyclotron Resonance Fourier Transform Mass Spectrometers (ICR-FT-MS) are unrivalled in terms of mass accuracy, precision and resolution in broad band scan.

Other than in TOF mass spectrometers, the m/z-time relationship is based on ion trajectory manipulation in a homogenous magnetic field. Once introduced into such a field, charged particles commence circular high frequency oscillation. Stronger magnetic fields cause oscillation at higher frequencies.

The circular oscillation is caused by the Lorentz force:

$$F = m\frac{dv}{dt} = zv \times B$$

where m is the mass, z is the charge, v is the velocity and the magnetic field strength is denoted by B. Rearrangements of this relationships lead to the mass to charge relationship

$$m/z = \frac{B}{w_c}$$

where w_c is the cyclotron frequency by which a given m/z oscillates [Marshall, A., et al., 1998]. Ion detection is possible by overlaying the circular oscillation of an ion with a radio frequency that matches w_c. In consequence, the given ion increases its oscillation radius and eventually comes into vicinity to detection plates, which are placed around the measurement chamber.

The resolution and mass accuracy of an ICR-FT mass spectrometer depends on the cyclotron frequency of an ion and on the duration of the oscillation, i.e. the time for which the oscillation can be detected. The free flight path in TOF mass spectrometers measures meters and the flight path in ICR-FT-MS measures kilometers.

By sweeping the excitation frequency over a range of frequencies it is possible to excite and analyze thousands of different m/z at the same time. This capability is important for metabolomics measurements because it enables the differentiation of isobars and different isotopologues. ICR-FT-MS measurements are not per se less quantitative than TOF measurements, however, since ICR-FT detection happens in a closed volume – the ICR cell – ICR-FT-MS is more vulnerable to oversaturation. Quantitative measurements of analytes require all ions to have minimum influence upon each other. Direct infusion injection into TOF-MS is just as non-quantitative as direct infusion into ICR-FT-MS. However, TOF mass spectrometers scan fast enough to support coupling to liquid chromatography. This technique separates the molecules and therefore minimizes interactions during ionization and detection. Still, TOF mass spectrometers do not provide sufficient resolution as to support good resolution of isobars or isotopologue peaks of different m/z species. ICR-FT-MS scans too slowly as to support LC-MS coupling.

In addition, liquid chromatography often requires strong pre-concentration of samples; in part because the ionization conditions for a wide range of analytes are not optimal in LC-MS coupling. Literature on EBC indicates such a strong dilution of metabolites that several milliliters of sample would have to be concentrated for the application of broad range LC-TOF-MS. Since, in addition, few is known about the metabolome of EBC, and since a principal goal of this thesis is to annotate as many analytes as possible, only ICR-FT-MS was used for the present manuscript.

1.2.4 Data Analysis

Metabolomics measurements produce large amounts of data, which have to be treated so as to obtain understandable and interpretable results. The data analytical techniques used for this purpose are data mining techniques and statistics – in that context often called chemometrics.

Data mining is used to extract potentially important data and statistics are used to verify its significance.

Data mining techniques are typically classified into unsupervised and supervised methods. Unsupervised methods – also called clustering algorithms – summarize data in a way that similar objects or variables are associated with each other (they cluster). Methods which pertain to this group are Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA) or K-means Clustering (K-means). The first two methods enable the clustering of data into its natural grouping and K-means clusters the data into K groups.

“Supervised methods” pertain to algorithms, which are first trained on a training set (commonly 1/3 of the data) to separate the data into a desired grouping. This training works by “fishing” variables that give a separation of the desired groups and joining them into a model. This group of variables is tested as to whether they separate the rest of the data as well and the goodness of the separation is verified by a statistic afterwards. If this statistic indicates, that the separation was significant (as well as specific and sensitive), the responsible variables can be said to represent the sample grouping and that they are therefore of importance for the experiment.

Supervised methods are for example Self Organizing Maps (SOMs) [von der Malsburg, Chr., 1973], Partial Least Squares or Projection on Latent Structures (both PLS) [Wold, S., et al., 2001], Support Vector Machines (SVMs) [Cortes, C., et al., 1995] or Random Forest Analysis (RF) [Breimann, L., 2001].

It is common praxis, to extract the important variables which either relate to a cluster of interest (usually one that separates a known grouping) or which relate to a successful supervised analysis outcome and to then perform statistics on them.

The number of literature references, which lay out the most commonly used data mining techniques is too vast to be cited here.

1.3 Metabolomics 2: Practical Aspects – Closing in on Reality

1.3.1 Practical Aspects of Instrumental Analysis

Where the copies of genes can be selectively amplified and detected, proteins can selectively be digested and thus sequenced, which allows for a direct comparison and matching of what we may call the DNA→RNA→Protein (DRP) trinity. Results on these levels can be directly associated to one and another. In the case of metabolites, however, relations to the DRP trinity are at best indirectly implied: in the case that a DRP includes a metabolite-specific enzyme or is regulated by a metabolite, it is possible to infer from function to sequence aspects and the other way around. In the case of lipids a much wider regulative cascade including many concertedly acting DRPs may lead to e.g. the general composition of a cell membrane. This composition may be regulated by surrounding cells or even much more distal tissues and organs. It may be regulated in part by proportions of transmembrane proteins, which form lipid rafts around them. Ultimately, membrane microstructures can be thought of as a basic recipe which is “hidden” on different genetic loci and the final membrane microstructure is a function of self-assembly. Additionally, the composition of the metabolome is – if at all – only vaguely predictable by means of DRPs as the total metabolite setup and its regulation depends on the cell-exterior supply with organic compounds and external regulation by e.g. the microbiome or environmental factors such as irradiation, mineral supply, temperature and so on. As the domains of DRPs are additionally a sink (or final destination) of metabolite fluxes – as they are polymers of metabolites – metabolites cannot be fragmented into sub-structures that correlate to the DRP domains. Metabolites cannot be amplified or over-expressed. At best genes of known relation to a metabolite can be knocked out by targeted mutagenesis or by insertion of genes, which produce the exactly mirrored RNA sequence as a function of the same promoter, which finally leads to an RNA knock out.

Ultimately, metabolites have to be quantified and identified solely at hand of their mass and their physico-chemical properties. This, however, is problematic for the following reasons:

- Metabolite concentrations vary from pico-molar (hormones) to molar scale (urea)
- Hydrophilicity can vary by 10 orders of magnitude: The substrates for the synthesis of

Sphingosine are Serine ($LogD_{pH=5.5} = -3.99$) and palmitic acid ($LogD_{pH=5.5} = 6.02$)

- Differences in pK_a, gas pressure, dipole moment and gas phase basicity vary in similar ranges as reported above

The analytical access to reaction partners can be largely impaired by strongly varying properties of metabolites. Additionally, given a constant chemical environment, one metabolite may entirely suppress another metabolite's response to a given analytical technique. These impairing factors concern with all techniques that are contemporarily used for metabolome analysis.

1.3.2 Practical Aspects of Data Analysis: Understanding the Methods

1.3.2.1 Prologue

Data Analysis pertains to two circles of methods, one of which – datamining – encompasses the extraction of important information from a given dataset and/or the creation of hypotheses. The other circle of methods – statistics – encompasses the verification of hypotheses, which either existed prior to instrumental analysis – hypothesis driven research – or which were created by means of data mining – data driven research.

In algorithms where each data mining iteration is first statistically verified before the next iteration starts, both circles are occasionally not distinguishable.

Data mining methods are often classified into supervised and non-supervised methods, where supervised methods are conditioned onto prior existing knowledge like a predefined classification of samples. Unsupervised methods develop a classification without prior knowledge. While this distinction is important for bio-informaticians and computer scientists, for the analytical chemist it is of less importance than the classification described below.

The second classification of methods is the differentiation between uni-variate methods and multi-variate methods. Univariate methods center on statistics on one variable over several objects at a time – it does mostly not encompass dataming steps. Clustering – a datamining technique – can be used in univariate analysis, but this is practically never done. Exceptions pertain to political sciences in the context of microaggregation. Multivariate methods encompass datamining steps and simultaneously treat multiple variables over several objects.

This classification of methods is – often unbeknownst – of large impact for the interpretation of results by biologists, analytical chemists, physicians or other end-users of the data analytical results.

The following sub-sections elucidate why this differentiation is important and where unconsciousness of the inherent differences can lead to confusions, especially in the context of surrogate marker definition.

Uni-variate and Multi-variate Methods

Apart from the fact that univariate methods address one variable at a time and multi-variate methods address sets of variables, it is important to ask how these methods work with the variables.

Uni-variate approaches are used in order to find out, whether the manifestation of a variable is significantly different (e.g. over-represented) in one sample set as opposed to another. Uni-variate methods can be understood as being “level approaches”.

As indicated above, multi-variate approaches fundamentally differ from uni-variate approaches. Independently of the algorithm used – be it PCA, PLS, HCA, SOM or correlation networks – multi-variate techniques are “relational approaches”. They are all based on types of similarity matrices; correlation-matrix, covariance-matrix, distance matrix, adjacency matrix (graph theory). These matrices are either computed prior to the actual classification (PCA, PLS, Networks) or are filled on the fly – while classification is performed (HCA, Random Forest).

Given a dataset with N variables, similarity matrices are squared N*N matrices, where each slot contains a scalar measure that describes the relationship between the variable in the respective row and the variable in the respective column.

The importance of the differentiation of both classes – “level approaches” versus “relational approaches” becomes evident when the typical workflow of published metabolomics papers is laid out schematically (Fig. 2).

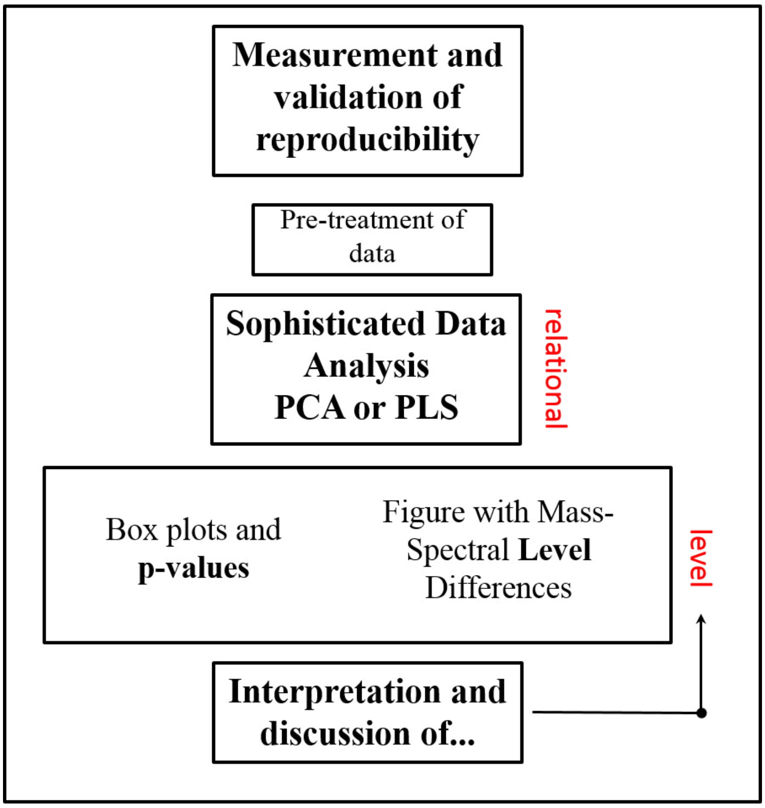

Measurement and validation of reproducibility
Pre-treatment of data
Sophisticated Data Analysis PCA or PLS
relational
Box plots and p-values
Figure with Mass-Spectral Level Differences
level
Interpretation and discussion of...

Figure 2: Schematic workflow of analysis in common metabolomics publications

It is common to perform “sophisticated” multivariate data analysis in order to extract important variables. Afterwards these important variables are tested for their level-differentiation and the variables that significantly differentiate in a uni-variate manner are reported. The fact, that multivariate analyses center on “relationship” and not on level differences commonly is unnoticed. The alleged “black box character” of multivariate tools is accepted and as a consequence a large amount of important information cannot be recognized, interpreted and much less published.

Black Box or no Black Box?

There are multivariate approaches, which also among specialists are considered as being black boxes and there are multivariate approaches which can be reasonably interpreted.

References in literature and the World Wide Web indicate that PLS and its variants as well as random forest clustering are indeed black box approaches, i.e. the cause as to why a variable supports data clustering cannot be reasonably explained. On OPLS-DA for example orthogonal signal correction is applied onto the data prior to PLS. As a consequence, data clusters in a Euclidean space which has a different basis or orientation than the original data itself. Results that are produced by such methods are sometimes not interpretable in a uni-variate fashion [practical experience]. Orthogonal signal correction is commonly applied when no valid model for the data separation can be devised by means of PCA or normal PLS. As will be discusses in chapter 3, the use of such “black box” techniques can be avoided by means of reasonable data pre-treatment.

Approaches which are no black boxes and are in their essence closely related to each other are PCA and correlation networking. Hierarchical clustering and K-means Clustering can be interpreted as well.

In order to support the intuitive understanding of results, only PCA and correlation networks are used throughout the present manuscript; univariate techniques are either assumed to be known or are introduced, where needed. The following sub-section first gives an overview of basic operations in linear algebra as well as some geometrical interpretations of these operations. Subsequently, PCA and graph theory – with a focus on correlation networks – are introduced and their relationship towards each other is elucidated.

1.3.2.2 Basic Operations of Linear Algebra as well as their Geometric Interpretation

Linear algebra pertains to the manipulation and analysis of vectors and matrices. Vectors are lists of numbers (scalars) and matrices are arrays of scalars. Mass spectra, which contain the variable pairs m/z|magnitude can be expressed as an N*2 matrix, where each row contains m/z in the first column and the respective magnitude in the second column. Basic vocabulary and operations of linear algebra are delineated below.

Scalar: a number; here a real number.

Vector: Vectors are lists of scalars. Vertical lists of scalars are **column vectors**, horizontal lists of scalars are **row vectors**. A vector in which **N** scalars are listed has **N** dimensions.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_N \end{bmatrix} \text{ Column Vector}$$

$$[x_1 \quad x_2 \quad \cdots \quad x_j \quad \cdots \quad x_M] \text{ Row Vector}$$

Vectors can be interpreted as lines in a Cartesian, N-dimensional space, which start at the origin. The vector entries remark the end coordinate of the line.

Magnitude of a Vector or the Euclidean Norm:

$$|a| = \sqrt{\sum_{i=1}^{N} a_i^2},$$

The Euclidean is the N-dimensional Pythagoras over a vector. A vector whose every element was divided by the Euclidean norm is a normalized vector of magnitude 1 (a unit vector).

Matrix: Matrices are rectangular arrays of scalars. Its rows are read as row vectors and are typically numbered from *i*=1:M; its columns are read as column vectors and are typically indexed as *j*=1:N. I a row vector is a matrix with N = 1 and a column vector is a matrix with M = 1.

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} & \cdots & x_{i,M} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,j} & \cdots & x_{N,M} \end{bmatrix}$$

N*M Matrix

Square Matrix: A matrix where N = M.

Transposed Matrix: A matrix rotated by 90°; rows become columns and columns become rows. Matrix X becomes Matrix X^T.

So Matrices are systems of vectors in the same coordinate space.

<u>Operations on Vectors:</u>

Addition of Vectors and/or Matrices:

$$\begin{bmatrix} 3 \\ -2 \\ 1 \end{bmatrix} + \begin{bmatrix} -2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 + (-2) \\ (-2) + 1 \\ 1 + 3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 4 \end{bmatrix}$$

Vector addition can be interpreted as „taking“ the line each vector represents and laying them end to peak without changing the direction of the summands. The above system can be interpreted as the summands being the cathetuses of a triangle and the sum being the hypotenuse.

Vector subtraction in turn can be interpreted as rotating the direction of the subtrahend vector by 180° and then laying its end onto the peak of the minuend vector. The difference between both is then the vector which connects the origin of the minuend with the peak of the subtrahend.

Scalar (Dot) Product

$$\begin{array}{c} [x_1 \quad x_2 \quad \dots \quad x_i \quad \dots \quad x_N] \\ \times \\ [y_1 \quad y_2 \quad \dots \quad y_i \quad \dots \quad y_N] \\ = \\ \sum [x_1 y_1 \quad x_2 y_2 \quad \dots \quad x_i y_i \quad \dots \quad x_N y_N] \end{array}$$

or

$$x \cdot y$$

Angle of two vectors x and y: $\cos(\varphi) = \frac{x \cdot y}{|x| * |y|}$

The angle of two centered unit vectors is likewise their (Pearson) correlation coefficient. This is an important basis for multi-variate data analysis.

The inner product of a matrix: Two matrices X and Y, which are to be multiplied have to have the exact same dimensions. Inner product computation works by first transposing Y into Y^T so that the row vectors of all N dimensions X compare to the column vectors of all M dimensions in Y^T. Then each matrix slot is filled with the scalar product of the incident row in X and the incident column in Y^T.

$$\begin{matrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \\ x_{i,1} & x_{i,2} \end{matrix} * \begin{matrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \\ y_{3,1} & y_{3,2} \\ y_{i,1} & y_{i,2} \end{matrix} = \begin{array}{cccccc} & & y_{1,2} & y_{2,2} & y_{3,2} & y_{i,2} \\ & & y_{1,1} & y_{2,1} & y_{3,1} & y_{i,1} \\ x_{1,1} & x_{1,2} & x_{1:M} \cdot y_{1:M} & x_{1:M} \cdot y_{2:M} & x_{1:M} \cdot y_{3:M} & x_{1:M} \cdot y_{i:M} \\ x_{2,1} & x_{2,2} & x_{2:M} \cdot y_{1:M} & x_{2:M} \cdot y_{2:M} & x_{2:M} \cdot y_{3:M} & x_{2:M} \cdot y_{i:M} \\ x_{3,1} & x_{3,2} & x_{3:M} \cdot y_{1:M} & x_{3:M} \cdot y_{2:M} & x_{3:M} \cdot y_{3:M} & x_{3:M} \cdot y_{i:M} \\ x_{i,1} & x_{i,2} & x_{i:M} \cdot y_{1:M} & x_{i:M} \cdot y_{2:M} & x_{i:M} \cdot y_{3:M} & x_{i:M} \cdot y_{i:M} \end{array}$$

Note that the indices for the transposed Y matrix were kept in order to indicate their original coordinate. The shorter side of the matrices is the M dimension and the longer side is the N dimension. The product XY^T of the matrices X and Y is an N*N matrix filled with dot products of the vectors along the M dimension.

In order to understand the importance of matrix multiplication, recall that many multivariate methods are based on operations on correlation or covariance matrices:

If the vectors in X and Y are normalized (and optionally centered) along M, the matrix XY^T is immediately the correlation matrix and without normalization XY^T is the covariance matrix!

This means that both, correlation matrix (CM) and covariance matrix of a data matrix D with N variables and M samples are directly accessible by calculating the inner product of the data matrix.

- If CM = DD^T, the correlation or covariance matrix pertains to the variables.
- If CM = D^TD, the correlation or covariance matrix pertains to the samples.

Eigenvectors and Eigenvalues: Eigenvectors and eigenvalues can only be created on a square matrix, e.g. a CM. An eigenvector **e** is a vector that, when multiplied with its respective square matrix **S**, results in a vector **v** that is a multiple of the eigenvector itself. The resulting vector **v** is either stretched, contracted or points into the opposite direction, however, it is co-linear with the eigenvector. The factor by which **v** is a contracted, stretched or inverted multiple of **e** is the eigenvalue. A multiple of an eigenvector is also an eigenvector. Likewise, an eigenvector on **S** is always an eigenvector on scalar multiples of **S**.

A square matrix of N dimensions usually has N eigenvectors, which each are associated with an eigenvalue. The set of all eigenvalues is denoted as the spectrum of the square matrix **S**. The product of all eigenvalues is the determinant of the matrix **S**, which in turn is associated to the volume **S** encompasses in N dimensional space. As a consequence, the relative proportion that the eigenvalue of an eigenvector has in respect to the sum over the spectrum of **S** (the trace) is associated with the proportion of variability that this eigenvector covers. In fact it is reasonable to say, that an eigenvector with a large eigenvalue covers a large part of variability within **S**. Eigenvectors of different eigenvalue are orthogonal.

Knowing these basic concepts it is now possible to explain and understand

- what PCA does.
- how different types of data pre-treatment affect the result of PCA.
- why it is not always appropriate to focus on uni-variate level difference when using multi-variate techniques.

1.3.2.3 Principal Component Analysis

Common schematic representations of what PCA is and how it works are the following:

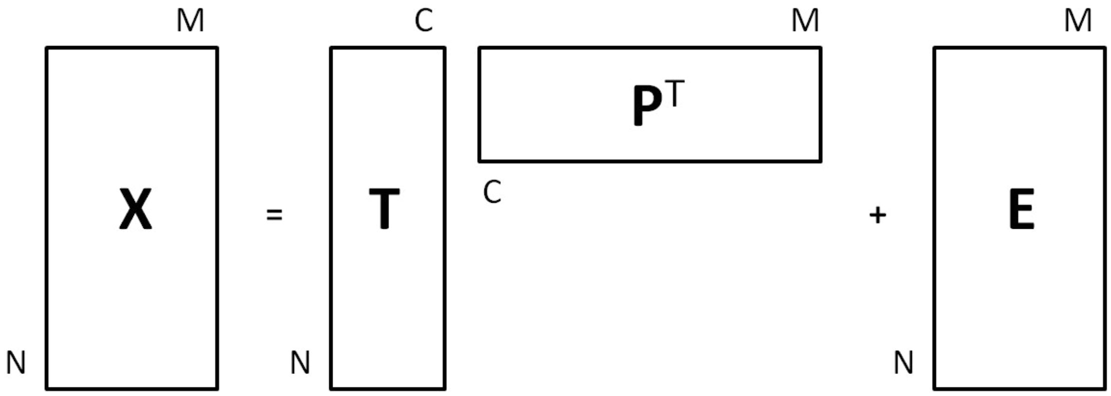


Figure 3: Schematic representation of PCA with X being the data matrix, T being a set of so called t-scores, P^T being the transpose of a set of principal components (eigenvectors) and E being the residual between a prediction $\dot{X}$ and X, where $\dot{X}$ is TP^T. $\dot{X}$ does often not occur in such schemes (because some knowledge is implied here).

Another representation which is commonly [whenever searching for PCA in the internet] presented in order to support understanding:

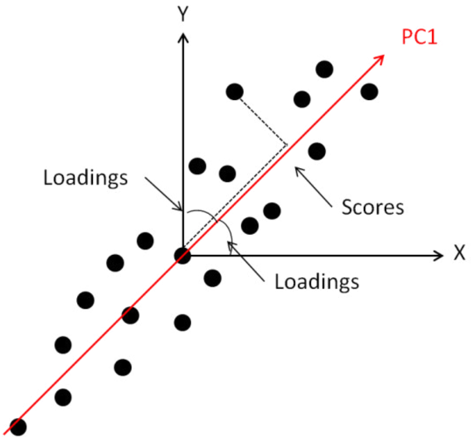


Figure 4: Second scheme supporting the understanding of PCA

The pre-omics era was characterized by the data consisting of more samples than variables, for which reason all classical representations of techniques associate the N-dimension of figure 3 with the samples and the M dimension with the variables. Due to the dimensionality of omics data, this notion has switched places.

In these schemes two important steps are often omitted. The first is as to how $\mathbf{P}^\mathbf{T}$ was generated and the second is as to how **T** (the scores which associate to the samples) were created. Commonly the first issue is completely omitted and the second issue is contextualized as follows: “**P** represents a new coordinate system and **T** represents the coordinates of the old data points in the new coordinate system.” [generalization].

Several misunderstandings – and misinterpretations – are induced due to these often metaphoric notions:

P is not directly related to the data entries in **X** because the set **P** contains the eigenvalues that pertain to the correlation/covariance matrix (**CM**) describing the relations between the samples; $\mathbf{CM} = \mathbf{X}^{\mathbf{T}}\mathbf{X} \rightarrow \mathbf{P}$ and not $\mathbf{X} \rightarrow \mathbf{P}$. Yet, the coordinates of **P** refer to the same coordinates along the M dimension. This means the magnitude of an entry in an eigenvector describes a large involvement of the corresponding variable into the correlation behavior over the samples that this eigenvector describes. Additionally, if **CM** is a covariance matrix, a large entry in the eigenvector indicates that the respective variable is either of large magnitude (and determinant for the interrelation of samples) or that it has moderate intensity but is associated with many other variables of the same trend and that they all are determinant for the interrelation of the samples.

The scores in **T** are calculated by $\mathbf{T} = \mathbf{XP}$ and refer to each sample; the entries of **T** are the scalar products of the “list of variables” and their impacts in P. Accordingly, a score describes, whether the magnitudes of the variables in the respective sample are in a relation to their impact on correlation structure. Conceivably, the metaphoric notions often used for the description of PCA are not helpful for the understanding of the technique.

The estimation $\dot{\mathbf{X}}$ of **X** is calculated by the term $\mathbf{TP}^{\mathbf{T}}$ in figure 3. Since the **T**-scores are scalar representations of the involvement of the sample’s variable set with the correlation structure, and overall there are as many eigenvectors as there are variables, the scalar entries in $\dot{\mathbf{X}}$ will at some point converge with the relative structure of **X**. Coherently, the difference $\mathbf{E} = \mathbf{X}\text{-}\dot{\mathbf{X}}$ converges to a minimum. The more eigenvectors are needed for **E** to be minimized, the more differently covarying variable groups are in the data. They all describe a specific part of the data; experimental impact and different sources of systemic error. Random error will never have large entries in PCA because they have poor correlation structure. A note at the side: It is now clear, that the representation in figure 4 does as well not show, what happens in PCA.

One often noted term in PCA is yet to be explained: the Loadings L, which refer to the importance of the variables. Loadings are commonly described as being the angles between the variables and the principal components (eigenvectors).

The formula L = **XT** is the only solution which allows a matrix multiplication along N and would therefore describe the variables. This again means that loadings cannot be the angles between the old variables and the principle components, as they do not have the same dimension. The loadings can only be the angle of each variable into the scores, so there is an indirect relation between the new coordinate system and the scores.

PCA is an unsupervised technique, which finds variable sets that correlate or co-vary given the data X. PLS (partial least squares or projection on latent structures) is a technique which basically tries to find the eigenvectors P that co-vary according to some grouping. It searches for the correlation/covariance structure in the data that reflects the grouping. These structures are already hidden in the PCA results and PLS supposedly finds them. PCA and PLS are extensively used approaches in metabolomics (not in other omics sciences!), however, in this thesis only PCA will be used.

After the true nature of PCA has been laid out, it is possible to present and discuss different problems, which underlie metabolomics analysis and to elucidate how different techniques of data pre-treatment help minimizing these problems.

1.3.2.4 Data Pre-treatment

Data pre-treatment is a matter that is largely under-addressed in metabolomics literature. It is a topic mostly discussed among bio-informaticians but it should be given much higher importance in the portfolio of anyone – also instrumental analysts – performing metabolomics. This is because bio-informaticians and statisticians are often not familiar with the confounding factors specific to an analytical aperture. Therefore, common techniques for data-pretreatment are merely reviewed briefly in this section. An in depth analysis of different techniques follows in chapter 3.

Data analytical techniques are bound to mathematical axiomism and therefore they require “cleaned data” in order to work properly. Data that is to be used for PCA or PLS analysis has to be multivariately normal distributed. This means that the entries in the correlation or co-variance matrix have to adhere to a normal distribution. Therefore, the intensity distribution over every sample and over the entire data matrix has to follow a normal distribution.

Three steps can be necessary to induce normal distribution of the CM: transformation, scaling/normalization and centering.

Remembering, that the CM is based on scalar products throughout the data, it is clear that the distribution of scalar products cannot be normal if the factors for this operation are not normal. Depending on the value of the signal to noise ratio in mass spectrometric analysis, intensity distributions of mass spectra (samples) are either power distributions or log-normal distributions with a power tail but always positive infinite. There is no negative value and there is a majority of peaks at small magnitude and a minority of peaks with a large magnitude [Lu. T., et al., 2005].

This means, the data has to be transformed so that the mean intensity of all peaks coincides with the median intensity of all peaks, i.e. the intensities need to follow a normal distribution. This is normally achieved by performing a log-transformation of the intensity entries. Transformation has to always be performed prior to normalization/scaling and centering. For log transformation, zero-entries have to be replaced by ones (they yield zero after transformation) and the basis of the logarithm has to be chosen to not be too large, because that would reduce the variation in the data too much.

After transformation, normalization/scaling is to be applied. Sometimes data transformation is omitted because some normalization/scaling techniques are believed to yield similar results.

Scaling implies, that intensity distributions over samples are already normal throughout all variables, but in comparison they are either multiplied by some factor or shifted by some value. Scaling is assumed to pertain to the variables, not the samples [van den Berg., R.A., et al., 2006]. Normalization pertains to spectra, but in their essence they are the same processes. The equations used for this procedure are the same, just that one is applied on the variables and one is applied on the samples.

Van den Berg et al. list multiple scaling techniques, which all imply parallel centering and “scaling” upon some value which is representative for the respective variable or observation. Here an inherent mistake can already be identified: Centering is performed using the un-scaled/un-normalized data. In terms of estimative statistics (mean, standard deviation, variance) this can be fatal, in terms of robust statistics (Median and inter-quantile ranges) this is not a major problem, as long as both compared entities have the same amount of non-zero values.

In general, first scaling should be applied, then normalization and then re-scaling. Centering should be applied afterwards. Details are discussed in chapter 3. In that sense, normalization is often neglected and scaling is regarded to be of more importance.

As indicated in the previous section, the differentiation between supervised and unsupervised methods is of minor importance than the differentiation between uni-variate and multi-variate techniques.

1.3.2.5 Network Analysis

The scientific discipline, which delivers the theoretical basis to network analysis, is called 'Graph Theory'. Graph theory started with Euler in the 18th century, as he tried to find a way of walking through Königsberg, trespassing each bridge over the Pregel River once. Despite being a very old scientific discipline it became recognized in omics science only recently. Rather than in life science, it was foremostly applied in computer science and sociology [Girvan, M. 2002]. Graphs are the mathematical expressions of networks. Graph theory is a multivariate technique for data analysis because it analyzes manifestations of pairwise relationships between variables. Graphs can be formulated on any pairwise relationship; be it physical interactions, probabilistic interactions or similarity, graphs can help formulating and analyzing issues not perceivable by statistics.

Graphs are mathematical models of the form G = (V, E) with V being a set of nodes of dimension N and E being a set of edges. While each v Є V can be associated to multiple elements of E, each e Є E can only be associated to a tuple of elements from V. Graphs are mathematical representations of networks and can be used to describe topological features of a network. An analyst who expresses his data in form of a network may want to know:

- Is there a path through the network, which touches each node only once?
- What is the least number of edges that have to be passed in order to get from node A to node B, i.e. what is the shortest path from A to B?
- How many shortest paths from A to B are there?
- Which other nodes are elements of this path?
- How many connections does each node have?
- How central is each node to the network, i.e. how important is it for the network structure?

The question of interest typically arises from the network's context, i.e. 'Is the graph directed or undirected?'; 'What was the criterion for the formation of an edge?'.

Road maps are an example of directed graphs where nodes are cross roads and edges are streets, while there are one way streets and two way streets. Undirected networks based on physical interaction are protein interaction networks. Edges in correlation networks are based on the similarity relation between node pairs, they are undirected as well. Metabolic pathways are directed graphs where edges reflect a reaction from substrate to product [Guimerà, R. 2005], they may be reversible (undirected) and irreversible (directed); one may see the equilibrium constants of reactions as a weight, which reflects the strength of directedness.

In linear algebra, graphs can be written as an N-dimensional square matrix of zeros in which each node is confronted to all other nodes of G and the existence of an edge is indicated by a non-zero entry (e.g. 1). This matrix **A** is called adjacency matrix and all properties of a given network can be calculated by means of linear algebraic operations on **A**.

An important measure in network analysis is the degree (connectivity) of a node. The degree is the number of connections a node is incident to. In the adjacency matrix, it is exactly the number of non-zero elements in the row vector that is associated to a given node (feature or variable). The higher the degree of a node is, the more other nodes it is associated to. If the degree distribution of a given network is not random, but adheres to a scale-free distribution or power distribution, then a node having a large degree is a rare and therefore significant occurrence. A network with such topology is as well said to be robust against random attacks but vulnerable to targeted attacks. That means, if a node of a scale free network is randomly selected and then deleted, the network structure stays intact. If a node is chosen and deleted because it has a high degree, the network structure will collapse.

Knowledge about the degree and degree distribution of a network is important in multiple types of networks; a correlation networks could for example pertain to genetic regulation. If a gene is causing the action of other genes, its knock out will impair cellular function dramatically. On the other hand, if an exon is knocked out (degree = 0), cellular function might be un-impaired.

Another network characteristic is the clustering coefficient. It indicates, whether the neighbors of a node are connected among each other as well. If all neighboring nodes are connected, they constitute a full graph, where all nodes are connected with all nodes. The number m of edges in a full graph of n nodes is known to be 0.5*n(n-1). A node of degree n must be associated to n-2 triangles. The number of triangles in an undirected graph is easily found by raising the adjacency matrix A to the power of three. The diagonal read outs indicate how many paths of length three exist from a node to itself. For metabolic pathways this is a rather unimportant measure; inspection of metabolic maps will rarely reveal triangles. A scenario, which would lead to a triangle between a, b and c would be if b-a = H_2, c-a = -H_2 and b-c = 2H_2.

Another measure that is often considered is betweenness centrality. It indicates how many of all existing shortest paths between all nodes run through a node of interest. Nodes of high betweenness centrality are necessarily important for questions of flux. In a metabolic pathway, the metabolite with highest betweenness centrality might be a very stable node, since it would have many supply routes.

There are many more network measures, but all of them are relatively useless if the network has a random topology [Barabási, A-L. 2004].

Network Clusters

Networks may have regions of high connectivity that are separated by regions of low connectivity. Members of such modules share more similarities with each other than they do with the rest of the network. Such modules are clusters and they can be found according to the Newman algorithm, which works on the basis of nodal degrees and eigenvectors [Newman, M. E. J. 2004a; Newman, M. E. J., 2004b]. The magnitude of eigenvector entries is proportional to the degree of the respective node. Grouping the nodes according to their eigenvector entries is the basis to the identification of network modules. A network can ultimately be seen as the graphical interpretation of PCA results.

1.4 Exhaled Breath Analysis

Breath analysis is a non-invasive technique for the diagnosis of possible pathologies. Already in ancient times the smell of the breath was a distinctive pattern to recognize diseases such as diabetes, liver, lung or renal pathologies or severe infections [Phillips, M. 1992]. Nowadays, breath analysis dresses a role in the detection of aging and neurodegenerative diseases and environmental pollutants or drug exposure [Risby, T. H., et al., 1999; Cao, W., et al., 2006]. The molecular composition of breath was characterized in healthy and pathological conditions, reporting as principal component (up to 99 %) nitrogen, oxygen, carbon dioxide, water vapor and inert gases [Miekisch, W., et al., 2004]. The remainder is composed of different kinds of molecules that range from parts per million concentrations to parts per trillions [Chen, S., et al., 1970; Pauling, L., et al., 1971; Riely, C. A. et al., 1974; Dannecker, J. R., et al., 1981; Solga, S. F., et al., 2010]. Volatile Organic compounds (VOC) are described to be present in normal subjects in a variety of 3400 molecules, constituted principally of isoprenes, alkanes, methylalkanes and benzene derivatives of which only a small partition is found in all screened subjects [Phillips, M., et al., 1999]. These compounds are an interesting target for the investigation of different pathologies [Risby, T. H., 2002] especially via high resolution analytical techniques [Risby, T. H., et al., 2006; Solga, S. F., et al., 2010].

Early studies on exhaled breath biomarkers are based on an analytical screening via GC

[Jansson, B. O., et al., 1969; Chen, S., et al., 1970; Pauling, L., et al., 1971; Riely, C. A., et al., 1974; Dannecker, J. R., et al., 1981] allowing the detection and identification of molecules with a concentration higher than 40µmol/mL. To increase the sensitivity in the detection, several sample concentration techniques, such as cryogenic trapping and adsorption onto carbonaceous or hydrophobic polymeric sorbents, were adapted. In order to detect a wider range of VOCs, especially oxidative stress markers, the analysis of methylated alkane was adapted. The oxidative stress modulates the DNA methylation levels [Campos A. C. E., et al., 2007]**.** Direct breath analysis is taken in account due to the minimized loss of sample. It is performed via electrochemistry, chemical sensors, optical spectroscopy, mass spectrometry, ion mobility, differential mobility spectroscopy, proton transfer mass spectroscopy or fast gas chromatography [Amann, A., et al., 2010].

Exhaled breath condensate analysis and its role in metabolomics

Exhaled breath condensate is a matrix constituted of three kinds of components: the droplets derived from aereosol formation from the airway lining fluid (ALF), the distilled water from the condensation of the water-saturated exhaled air and the water-soluble volatiles in the condensed breath. The studies on EBC allow ALF monitoring in health and disease [Hunt, J. F., 2002; Kharitonov, S. A., et al., 2001; Mutlu, G. M., et al., 2001]. Metabolomics on EBC focuses its interest onto the non-volatile compounds and the water-soluble fractions [Hunt, J., 2002]. The most prominent compounds described in scientific literature until 2012 (850 publications examined) are inorganic compounds (347 publications) such as hydrogen peroxide, nitric oxide and gaseous compounds. The second class of frequently described compounds encompasses isoprostanes/prostaglandins/prostanoids (114 publications) followed by leukotrienes (103 publications).

One of the main bottlenecks that concerns EBC sampling is the dilution factor [Effros, R.M., 2010]. The dilution is due to water vapor, which is exhaled as a product of metabolic processes, and which condenses in the cooled sampling process. The ALF compounds in EBC can be diluted in a range from 20 fold to 30 000 fold [Effros, R. M., 2010; Effros, R. M., et al., 2002]. Additionally, the inter day and subject variability in dilution needs to be considered. Commonly, internal standards are necessary in order to estimate the dilution factor of the ALF aerosol. One of the standards used is urea [Rennard, S. I., et al., 1986], because of its good diffusion, its proper distribution in all body compartments (renal papilla excluded) and non-excessive metabolization in the lungs. It is assumed that the dilution factor of non-volatile compounds can be estimated via urea concentration measurements (considering the interstitial urea as well). Different models, based on gender and age, are used for measuring blood urea nitrogen (BUN) [McPherson, K., et al., 1978] and estimate interstitial urea concentration. Examples of measured urea concentrations in EBC (with consequent dilution factor calculation) are reported in studies on protein, albumin and ammonia in EBC [Dwyer, T. M., 2004].

Another limiting factor in EBC research is the complexity in comparing different studies, due to the variations derived from different collection techniques [Horvath, I., et al., 2005] and from the lack of standards that can be used in EBC research [Davis, M. D., et al., 2012]. Studies comparing inter- and intra-individual variability of EBC biomarker measurements, due to different collection techniques are reported [Do, R., et al., 2008]. The response of long term sample storage, short term sample storage and sample volume were evaluated monitoring acid stress biomarkers, pH, and ammonia.

Metabolomics research on EBC is a topic that has been addressed by no more than a dozen publications [e.g. Bertini, I., et al., 2013; Sofia, M. et al., 2011; Montuschi, P., et al., 2012]. The only study that has included EBC into systemic metabolome screening is the HuMet study [Krug, S., et al., 2012]. For this reason there is no body of knowledge as to how the above mentioned complications in EBC analysis affect the mass spectrometry based deep screening of the metabolome.

1.5 Diabetes mellitus

Diabetes mellitus is one of the most globally spread chronic metabolic diseases, recorded as one of the five leading causes for death in developed countries [International Diabetes Federation, 2011]. Its main feature is a failure in liver metabolism, which leads to a malfunction of insulin and glucagon [Harris, M. et al., 1997]. In diabetes mellitus these two hormones are not able to maintain a constant blood glucose level. One of the key hormones in the glucose homeostasis, insulin, is being produced in the β-cells of the pancreas. Its role is important in the absorption of glucose from the blood by liver cells, skeletal muscle and fat tissue. In order to accomplish the absorption of glucose, it inhibits the release of glucagon. The latter is produced in the α-cells in the Langerhans Islets as answer to a decrease of blood glucose concentration. When the level of glucose remains high in the blood, the condition of hyperglycemia is present. There are three main classes of Diabetes mellitus: Type 1 Diabetes, Type 2 Diabetes and gestational Diabetes (GDM). The first form of Diabetes is caused by an autoimmune reaction against the pancreas cells, which causes a malfunction of the insulin secretion [Medvei, V.C., 1993]. The second form of Diabetes is the most common form world wide. In this case the liver is insulin resistant or a condition of impaired secretion is observed (in some cases both conditions were recorded). This form of the disease correlates with other metabolic disorders that belong to the metabolic syndrome, such as: high blood pressure, high cholesterol levels, high triglycerides, high inflammatory marker levels and central and visceral obesity. All these factors lead to a risk of heart disease and cardiovascular complications [International Diabetes Federation, 2011]. The monitoring of the glucose and insulin levels in the fasting state, allows the diagnosis of diabetes via ISI Matsuda index calculation [Matsuda, M., et al., 1999]. This index is calculated after subjecting patients to an oral glucose tolerance test (OGTT). The results lead to the observation of the whole body insulin sensitivity. Both types of Diabetes are correlating with lung dysfunctions [Goldman, M. D., 2003]. Investigation of EBC may be a useful tool to study different markers such as inflammatory markers in order to understand mechanisms related to diabetes and in order to enable diagnosis of early states of insulin resistance by means of non-invasive sampling.

1.6 Aim of the Thesis and Outline

1.6.1 Aim of the Thesis

The aim of the present thesis is to develop an analytical workflow that allows for the extraction of mass spectrometric surrogate marker candidates from exhaled breath condensate. Particularily, the final workflow is intended to enable the extraction of markers for diabetes mellitus from EBC. As there is currently no proof for the involvement of the EBC metabolome with systemic metabolism; the equivalent aim is to establish this link.

The workflow is intended to maximize the amount of metabolically relevant information in the context of non-targeted metabolomics. The term 'Deep Metabotyping', which is used in this manuscript's title is intended to emphasize this aim as opposed to the aim of identification and quantification in parallel.

The present manuscript is intended to pave the way for future applications of EBC beyond the scope of pulmonary diseases.

1.6.2 Outlining the Thesis

Chapter 2 will continue with the introduction, extension and evaluation of mass difference network based annotation (Netcalc). Central topics will be mass spectral calibration, the introduction of the Netcalc algorithm, the specification of elemental filters, an adaptation of the Netcalc transformation sets towards applications in metabolomics and the investigation of sources for false annotations.

Chapter 3 focuses on data cleaning, especially the control of binary dependencies, the power-nature of signal distributions, normalization techniques and their effects as well as the use of Netcalc in the elimination of co-linearity. Ultimately a network-based normalization workflow is introduced.

Chapter 4 uses a dataset on smokers and non-smokers in order demonstrate how mass difference networks, co-intensity matrices and their eigenvectors can be used to mine and understand different types of surrogate markers and matrix effects. In addition, mass difference enrichment analysis is introduced as a method that can support data interpretation in case of lacking database support.

Chapter 5 will use the approaches developed in Chapters 3 and 4 to analyze EBC data from the HuMet study. In this study, 15 volunteers were led through five nutritional and metabolic challenges: 36 hours of fasting (F), ingestion of a standard liquid diet (SLD), oral glucose tolerance test (OGTT), oral lipid tolerance test (OLTT) and a physical activity test (PAT). The study was designed for the investigation of the normal dynamic range of the human metabolome. Here, this study is used in order to establish the link between EBC and systemic metabolism, which in extension gives prospects towards the screening of diabetes mellitus.

Chapter 6 will summarize the thesis, evaluate to which extent the objectives of the thesis were met, and develop an overall workflow for ICR-FT-MS based metabotyping. An outlook towards future research is given.

2 Netcalc. A Network based Annotation Algorithm and its Adaptation to Metabolomics

Annotation of mass spectral peaks is the assignment of either a sum formula (putative annotation) or an identity (annotation or identification) to a mass spectrometric peak. In literature the term ‘annotation’ increasingly refers to ‘putative annotation’ and not to identification. In theory, metabolic profiling can be performed independently from metabolite annotation/identification, since a phenotype specific m/z-profile itself does not require such information. However, m/z feature annotation has a multitude of advantages:

1) Broad scan mass spectra contain a multitude of co-linear or redundant information, since multiple features can be thermal adducts with solvent molecules, other molecules, different charge states of the same feature, fourier transform artifacts or isotopic peaks. The presence of such co-linear features hampers data analysis as non-random co-linearity leverages the underlying correlation structure. Feature annotation and omission of non-annotated features can improve the data analytical situation.
2) Feature annotation supports the evaluation of sample processing aspects like the introduction of contaminations. To that end, feature annotation helps assessing whether a peak can be related to the sample or to extraneous factors.
3) For data to be analyzed, singular mass spectra have to be translated into a feature*sample matrix, which is only possible if data is adequately calibrated or if two features from different mass spectra can be identified to be the same with sufficient accuracy. The appearance of redundant annotations can give hints as to whether calibration and alignment processes were performed appropriately.
4) Early annotation of features helps estimating physico-chemical properties of features and enables the analyst to devise appropriate strategies for targeted analysis.
5) Feature annotation supports the formulation of (bio-)chemical hypotheses, which can then be tested in bioassays, on cell cultures, other model systems and/or proteomics/genomics databases. It helps understanding the processes underlying the investigated phenomenon and supports causative inferences. To that end, annotation helps to devise strategies for the manipulation/treatment of a phenomenon.

2.1 Prior to Annotation: Calibration and Error Distributions

Calibration is a process by which measurement errors are estimated in order to correct for them. The simplest model of the relation between theoretical variables and empirical variables can be expressed as:

$$X_{t,m/z} = kX_{e,m/z} + E_{m/z}$$

With X_t being the set of theoretical values, k being a factor, X_e being the set of empirical values and E being the set of errors; with all sets being a function of m/z. Calibration is performed by approximating k so as to minimize E. In praxis, E follows a close to Gaussian distribution – an error distribution – at any nominal m/z. Its mean is to be centered on zero at any given m/z.

Classically, calibration of mass spectra is performed at hand of the error distribution pertaining to a set of spiked standards, which optimally span the entire m/z range of the spectrum.

As discussed in chapter 1, standard spiking is to be avoided in direct infusion MS, since the standards may suppress the signals that are intrinsic to the sample. In default of standard reference peaks, mass spectral calibration is performed against a list of well known reference masses by listing the m/z values which are the closest to that of the reference masses and by then approximating the observed error distribution.

This internal calibration is based on the process of mass matching and is therefore vulnerable to a number of factors, which in the end lead to a nominally good, yet false approximation of E, ultimately leading to an offset of the true error distribution. As a consequence, it is possible that annotation results appear to be accurate where they are not.

Scenarios in which a calibration offset can be observed are:

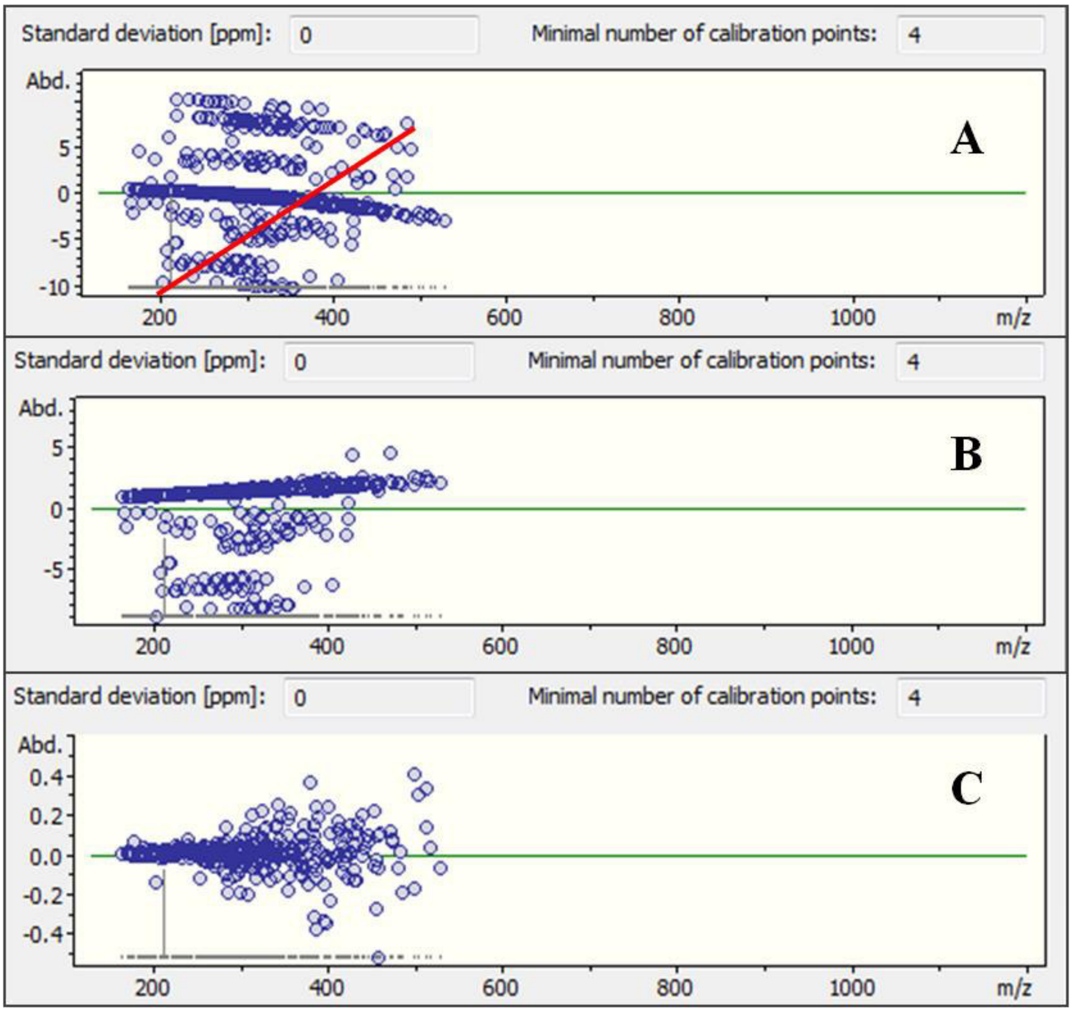
- Choice of the wrong function for error approximation; e.g linear approximation when the distribution is in fact of higher order
- Choice of a too small set of reference masses, and therefore failure to approximate the error centers
- Choice of a reference mass set, which contains isobars instead of the actual identities of the empirical peaks (this is often dependent on the sample type)

Types of error distributions depend on the mass spectrometer. ICR-FT-MS distributions are close to linear but depending on ion density slight higher order deviations at the beginning and end of the m/z dimension may occur (Horwitz trumpet). TOF mass spectrometers tend to have polynomial error distributions; the number of polynomials can range from 2 to 6. The higher the number of polynomials used, the higher are the degrees of freedom of the approximation (i.e. there are multiple minima of E). In consequence, the more polynomials are used for approximation, the more data points or reference mass hits have to be available in order to avoid over-fitting.

High degree polynomial error distributions have the disadvantage, that they do not vary linearly but polynomially, which means that the displacement of the error distributions' center is not constant over the m/z range and time; resulting local offsets of the error distribution are difficult to identify if the attempt to identify them is undertaken at all. Softwares of mass spectrometer vendors offer to control this problem by means of lock mass injection but the corrections for variations of these constantly injected reference analytes are typically linear which leads to calibration offset without the user noticing it. In such cases, an annotation error reported to be close to 0 ppm is void of essence.

Calibration is not only important for the correct annotation of single mass spectra, it is essential for the unification of single mass spectra into an m/z*sample matrix (sample matrix), which is needed for data analysis. If all spectra that are to be unified into such a matrix are calibrated in the same way, i.e. their error distributions are the same, then they will have good alignment. In this case, annotation of the unified m/z list gives a well shaped error distribution (Gaussian with constant standard deviation for all m/z bins). However, if the calibration was not uniform throughout all spectra, the distribution will be dispersed and consequent annotation results cannot be trusted.

The calibration strategy adopted in the present manuscript relies on the annotation of the mass spectrum with the most m/z peaks by means of Netcalc. An invaluable advantage of the annotation strategy that was introduced by Tziotis [Tziotis. D, 2011] is that the annotation process 'walks along' the error distribution. In consequence, an annotation run of a raw spectrum can be used to give the best possible approximation of a 'taylor made' reference mass list. Such a large reference list, which may contain hundreds of references, allows for a good resolution of the error distribution.



Standard deviation [ppm]: 0
Minimal number of calibration points: 4

Standard deviation [ppm]: 0
Minimal number of calibration points: 4

Standard deviation [ppm]: 0
Minimal number of calibration points: 4

Figure 5: Screen shots of the calibration process using a reference spectrum that was pre-annotated by Netcalc

Figure 5 A shows multiple error distributions and one large, dense and centered distribution. Assuming this distribution to be the appropriate reference, we crop the variables at the upper end (5 B) and the lower end (5 C). The result is a well centered and dense error distribution, which leaves any doubt concerning alignment behind. The red line in 5 A indicates just one of many calibration results that might occur using insufficient numbers of potentially inappropriate reference masses.

Abnormal error distributions can be recognized and omitted from the data set. It is difficult to even discover abnormalities if calibration is performed with four points only. The detection of such abnormalities is crucial for spectral alignment. It is probably a rare phenomenon, but in the preparation for the present manuscript, one slightly mis-calibrated mass spectrum was sufficient to cause mis-alignment of more than 200 accurately calibrated mass spectra.

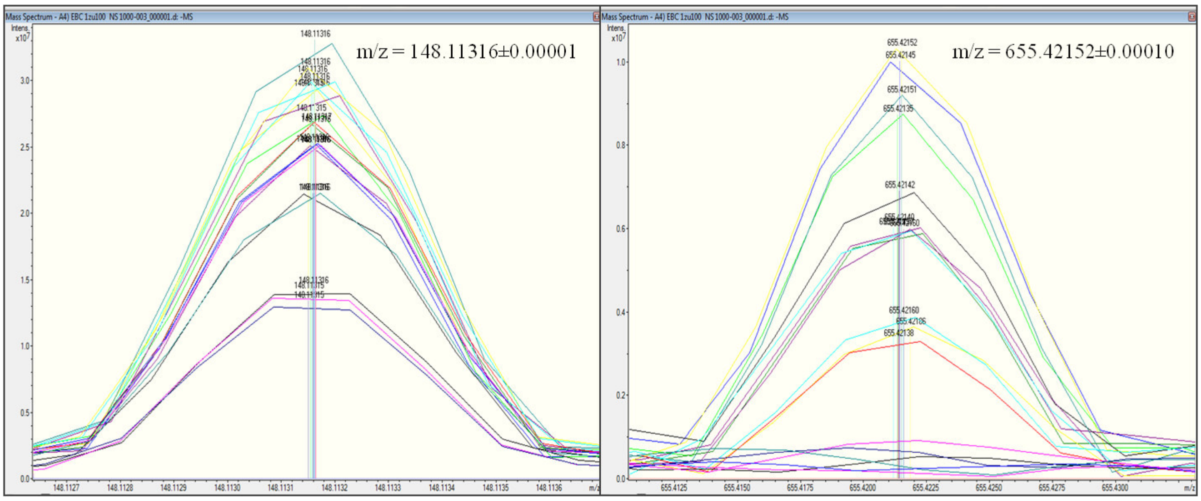


Figure 6: Alginment of 15 mass spectra at the lower and upper end of the acquired m/z range. The m/z deviations are 0.07 ppm at m/z = 148 and 0.15 ppm at m/z = 655. Using the above introduced workflow the calibration of each spectrum took more or less 30 seconds.

Perfect alignment was possible only after identification and omission of the mis-calibrated spectrum. It is therefore more important to verify zero alignment of the spectra, than to find a constellation of reference masses that provides an as small as possible standard deviation of E. The first strategy in order to obtain a sample matrix consisting of trustworthy annotations is to annotate each spectrum, to then omit all non-annotated m/z values and then to unify the exact theoretical masses of the respective annotations.

If the mass spectra are well aligned then the second (and more time saving) strategy is to unify all spectra before annotation. This second option can turn out to be more reliable as there are different annotation strategies, which all have their inherent advantages and disadvantages.

2.2 Annotation Strategies

There are two traditional approaches and one novel approach to m/z feature annotation.

The first traditional approach – the combinatorial approach – involves the calculation of elemental combinations, which as close as possible add up to the same mass as the given m/z (z = 1) value with consequent identification of isotopologues and comparison of relative theoretical isotopologue abundances versus the observed abundances.

The second technique can be called database matching. In this case m/z signals are queried against a metabolite-m/z list stored in a database.

The third annotation strategy, introduced in 2011 by Tziotis et al. is a mass difference network based algorithm called Netcalc. M/z differences between mass spectral peaks, acquired on a high accuracy mass spectrometer, can be associated to a specific elemental difference between these peaks. For example: Two peaks which have an m/z difference of 14.015650 m/z are related to another by a sum formula difference of [CH_2]. Consequently, if the smaller mass has the formula $C_2H_6O_2$, the larger mass must have the formula $C_3H_8O_2$. This principle is used in tandem MS with the mass of a neutral loss being representative for the loss of a certain combination of elements in response to fragmentation of a molecule. Netcalc searches for all pairs in a mass spectrum, which can be brought into relation by a list of Δm/z|Δelement relationships (later called REMDs for *Reaction Equivalent Mass Differences*) and builds up a network from them. Finally, the (hypothesized) knowledge of the formula of only one member (node) of the network implies the knowledge of all other nodes' formulae that are reachable through assigned network edges (Δm/z|Δelement relationships).

In its essence, Netcalc uses cross linked homologous series and it was first applied in order to extract and visualize exactly these homologous series from NOM data. For this reason elemental REMDs were used, which are not optimal for the description of multi-step biochemical reactions as they occur in metabolism.

In the following sections the Netcalc algorithm is explaind in depth, elemental filters for the exclusion of false annotations are introduced and an REMD list which is adapted to metabolomics data annotation is developed. Afterwards, extensions to the algorithm which minimize the occurrence of false annotations are introduced and the performance of Netcalc in respect to the combinatorial approach and the database matching approach is evaluated. This evaluation centers on: the proportion of annotated data to non-annotated data, false positive annotations, error distributions, annotation of noise perturbed data and finally the identification of error sources.

2.3 Netcalc

Metabolic networks can be generalized into stoichiometric networks in which each edge describes the stoichiometric change that occurs during a (bio-)chemical reaction. The use of ultra high accuracy mass spectrometry allows for the replacement of these stoichiometric changes by differences in molecular ion mass. The resulting mass difference networks enable the researcher to translate mass spectra into stoichiometric networks, which are the immediate link to the description of metabolic networks or metabolic pathways [Breitling, R., et al., 2006].

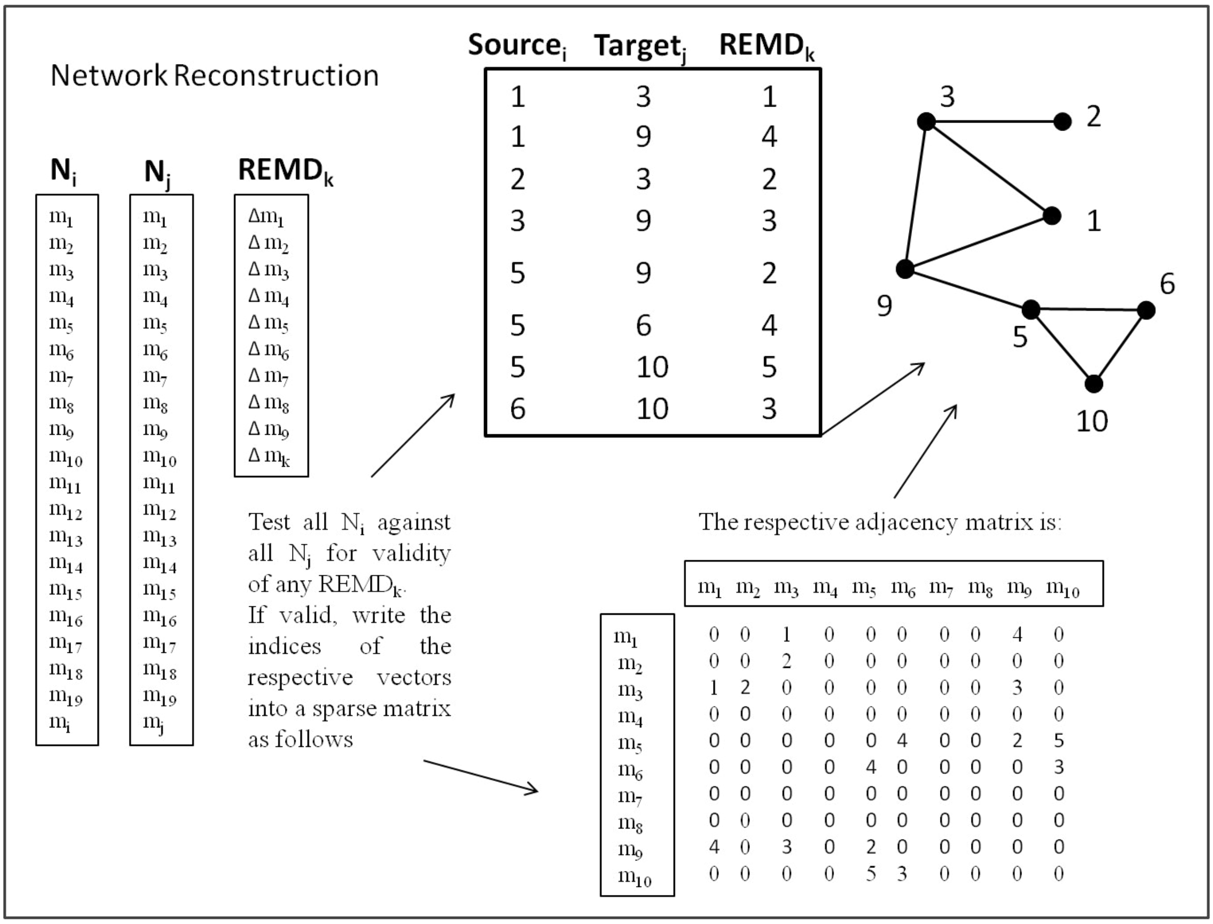
Mass difference networks have a useful property which was left unattended by Breitling: Each node which is adjacent to another node defines the neighboring node's molecular formula by virtue of the stoichiometric transformation described by the incident edge. Consequently, Tziotis et al. used this property and defined that the molecular formula of each node in G is immanently accessible given the molecular formula of only one other accessible node, if G is a graph component in which no unconnected node exists. By this virtue they defined the Netcalc algorithm which has a wide spectrum of applications for metabotyping.

2.3.1 The Concept

The first step of the Netcalc process is the construction of a network by comparing each ith and jth entry as to whether they match one of k reaction equivalent mass differences (REMDs). The equivalence criterion is the edge formation error (EFE), which is calculated as follows:

$EFE = 1000000*||m_j-m_i|-REMD|/(0.5*(m_j-m_i))$.

It is the absolute deviation of the absolute difference between the masses m_i and m_j from a given REMD, expressed over the mean mass of both 'reaction partners' in ppm.


Network Reconstruction
N_i
m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_i
N_j
m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_j
$REMD_k$
Δm_1 Δm_2 Δm_3 Δm_4 Δm_5 Δm_6 Δm_7 Δm_8 Δm_9 Δm_k
3
2
1
9
6
5
10

Test all N_i against all N_j for validity of any $REMD_k$. If valid, write the indices of the respective vectors into a sparse matrix as follows

$Source_i$	$Target_j$	$REMD_k$
1	3	1
1	9	4
2	3	2
3	9	3
5	9	2
5	6	4
5	10	5
6	10	3

The respective adjacency matrix is:

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
m_1	0	0	1	0	0	0	0	0	4	0
m_2	0	0	2	0	0	0	0	0	0	0
m_3	1	2	0	0	0	0	0	0	3	0
m_4	0	0	0	0	0	0	0	0	0	0
m_5	0	0	0	0	0	4	0	0	2	5
m_6	0	0	0	0	4	0	0	0	0	3
m_7	0	0	0	0	0	0	0	0	0	0
m_8	0	0	0	0	0	0	0	0	0	0
m_9	4	0	3	0	2	0	0	0	0	0
m_{10}	0	0	0	0	5	3	0	0	0	0

Figure 7: Construction of mass difference networks.

The lower right matrix in the figure is a so called adjacency matrix, which assigns the ID of each found REMD to pairs of features m. In order to save the computational space and complexity, the adjacency matrix can be re-written as a sparse matrix (upper matrix in figure 7). Both matrices are a blue print for the network given in figure 7.

If we now assign a formula to node 2, the formula of all other nodes can be calculated by walking along the edges. In theory the degree of the starting node is not of importance, for which reason it does not matter, whether we know the formula of node 2 and start from there, or whether we know the formula of node 9 and start from there. However, in praxis, and by experience, nodes with low degree tend to be associated with wrong edge assignments. Nodes of higher degree necessarily have many partners, which validate their involvement into the chemistry that is represented by the network. Nodes with low degree consequently represent rather exotic formulae, which have not more than one potential chemical relationship in the network. Reasons for low degrees are either based on the true chemical context of the mass spectrum, or they are distal to the spectral error distribution, or they have a false edge assigned. The reason for false edge assignment will be investigated in chapter 2.6.

2.4 Elemental Filters

Elemental filters narrow down the solution space of elemental combinations by testing combinatorial solutions for their validity in terms of electron configurations. By watching at the electron configurations of elements it is possible to tell how many covalent bonds an element can form. The electron configurations of CHONS and P are shown in table 1.

Table 1: Electron configurations of CHONS and P.

Element	1s	2s	2p	3s	3p	3d	Number of Valence Electrons (VEs) and bonds
Carbon	$1s^2$	$2s^2$	$2p^2$				$2s^2+2p^2$ = 4; 8-4 = 4 bonds
Hydrogen	$1s^1$						$2s^1$ = 1; 2-1 = 1 bond
Nitrogen	$1s^2$	$2s^2$	$2p^3$				$2s^2+2p^3$ = 5; 8-5 = 3 bonds
Oxygen	$1s^2$	$2s^2$	$2p^4$				$2s^2+2p^4$ = 6; 8-6 = 2 bonds
Sulfur	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^4$		$3s^2+3p^4$ = 6; 8-6 = 2 bonds*
Phosphorus	$1s^2$	$2s^2$	$2p^6$	$3s^2$	$3p^3$		$3s^2+3p^3$ = 5; 8-5 = 3 bonds*

*Sulfur and phosphorus can delocate (promote) their paired electrons into the d orbitals. S, p and d orbitals then hybridize to give an sp^3d orbital for phosphorus (allowing for five bonds) and an sp^3d^2 orbital for sulfur (allowing for six bonds).

Traditional elemental filters are based on elemental ratios that are allowed for different elements. For example: van Krevelen diagrams are typically based on a range of H/C ratios ($0.5 < H/C < 2.5$) and O/C ratios ($0 <= O/C <= 1$) [Hertkorn, N., et al., 2008]. Expressing filters in terms of such ratios is convenient but in case of metabolites they may be inappropriate (phosphorylated compounds can have $O/C > 1$).

It is therefore a more accurate solution to build up a filtering algorithm, which is based on the principles of covalent bonding; on electron configurations. Such a filter is elaborated in the subsequent section.

2.4.1 The Seven Golden Rules

The golden standard paper for m/z formula annotation is momentarily considered to be the work of T. Kind and O. Fiehn, [Kind, T. and Fiehn., O., 2007] which state the seven golden rules for formula annotation.

The "golden rules" can be separated into two groups: 1) rules that limit the degrees of freedom and 2) rules to validate candidate formulae.

Group one constitutes of rule #1 and rule #6. Rule #1 suggests a database driven limitation of maximal element counts throughout different mass ranges. It is necessary for purely combinatorial annotation algorithms in order to minimize the solution space. In principal, rule #6 does the same as it suggests a database-driven filtering of maximally observable co-occurrence of elements. This rule restricts the solution space by excluding rarely observed combinations of elements. These rules are relatively unnecessary in Netcalc annotation, since the number of elements is restricted by the number of elements in each reaction and the maximum number of times each reaction can occur as a homologous series within the observed mass range.

Group number two constitutes rules #2, #3, #4 and #5. Rule #7 does not concern with this thesis as it centers on GC-MS.

Rule #3 enforces the necessity to validate molecular formulae by means of isotopologue ratios. Without a doubt only isotopologue patterns can validate molecular formulae at any mass spectrometric resolution. However, a molecule which is meant to be validated by this measure needs to occur in such high abundance, that only targeted analysis can guarantee the detection of the isotopologue micro-structure which is necessary for that feat. The fewest chemical analysts are aware of the fact, that nominally equal isotopologues of different elements exhibit different mass defects, i.e. that they have different mass. A verification especially of complex analytes containing more than just CHO can only be performed with ultra-high resolution (R> 100 000). At lower resolution it is not possible to distinguish isotopologues caused by different elements; their mass spectrometric peaks tend to merge which causes a shift in the averaged, non-resolved isotopologue peak. In targeted settings it is very well possible to achieve sufficient ion abundances and (mixed) isotopologue peaks which are still specific for a molecular formula.

The common scenario in non-targeted mass spectrometry is different:

The $^{12}C/^{13}C$ ratio of glucose is 6.718 at R = 90 000. Given an S/N marigin of 3, the minimum S/N that the 12C peak has to satisfy is $S/N^{13C}_{min,\ 12C} = 3*100/6.718 = 44.66$. Analogously, $^{18}O/^{12}C = 1.254$ at a margin of S/N = 3 gives $S/N^{18O}_{min,\ 12C} = 3*100/1.254 = 239.23$. Consequently, the identification of the sum formula $C_6H_{12}O_6$ by means of isotopologue abundances requires a glucose concentration, which is high enough to cause an $S/N \geq 239.23$.

Different mass spectrometers allow for different S/N margins. ICR-FT-MS allows for a very accurate calculation of the local noise levels due to the large numbers of picture points per spectrum (1M or 2M). TOF instruments develop a less constant noise pattern which necessitates higher S/N margins like S/N > 100, which results in

$S/N^{18O}_{min,\ 12C} = 100*100/1.254 = 7974.48$.

Only a small proportion (1% to 5%) of ^{12}C peaks fulfill these requirements, which makes isotopologue based formula validation a limiting task given the aim of metabolomics, to maximize the range of metabolite detection.

Performing non-targeted metabolomics and following the dogma to maximize the content of information carried by a mass spectrum, it is necessary to omit rule #3 while being in the non-targeted phase of the metabolomics workflow.

Rules #4 and #5 treat valid ranges of elemental ratios. Elemental ratios are a common tool in the analysis of complex mixtures like natural organic matter, which are often thermodynamic conversion points in structural chemistry. For this reason marginal elemental ratios are often $0.5 < H/C < 2$ and $0 < O/C < 1$. The metabolome, however, constitutes of many exceptions to these largely accepted margins. An obvious way to define margins for elemental ratios is to investigate large and representative databases of chemical compounds, which is exactly the way how rules #4 and #5 were created. However, all valid ranges of elemental ratios must be justifiable by the rules of chemical bonding, which makes rule #2 a super rule to elemental ratios.

Rule #2 of the seven golden rules refers to the valency of the elements which make up a neutral molecule (m/z values refer to ions, which first have to get neutralized by correcting for

1.007276 Da in the case of (de)protonation events). Here the octet rule of LEWIS was extended by the rules of SENIOR [Senior, J. K., 1951], which are relatively unknown to chemists (despite the fact that they are the mathematical origin of the degree of unsaturation). The SENIOR rules as stated in Kind [Kind, T. and Fiehn, O., 2007] and Morikawa [Morikawa, T. and Newbold, B. T., 2003] are read as follows:

1) The sum of valencies is an even number, or the total number of atoms having odd valencies is even.
2) The sum of valencies is greater than or equal to twice the maximum valency.
3) The sum of valencies is greater than or equal to twice the number of atoms minus 1.

An inspection of these rules shows that rule 3) is different in that its minimum criterion implies an odd sum of valencies, while the other rules state an even sum of valencies to be obligatory. Rule 3) also violates the other statements regarding the odd-even parity in Morikawa et al. These violations are in fact the result of the verbal formulation of rule 3). Rule 3) is better written as “half the sum of valencies is grater than or equal to the number of atoms minus 1”. While rule 3) was correctly implemented in the software of Kind and Fiehn, the verbal description of Morikawa does not reflect Senior’s original emphasis. Verbal descriptions of equations should allways be accompanied by the equation itself in order to avoid misunderstandings. Further investigation of the original emphases of SENIOR 1951 will line out a wider spectrum of applications of his work.

2.4.2 Senior’s Rules, the Cyclomatic Number and Extended Hybridization

Molecules of interest for metabolomics investigations are composed of the elements CHONS and P. The valency of these atoms indicates how many bonds they can form with other atoms. The valencies for our elements are: C(4), H(1), O(2), N(3), S(2,4,6) and P(3,5). Now what are molecules? Essentially, molecules are networks (or graphs) that are composed of elemental atoms (nodes) and covalent bonds (edges) and each element is known to engage in the number of bonds that correspond to their valency. The valency is the degree of a node in an atomic network. The sum of all valencies in networks is exactly twice the number of edges (bonds).

An example: Acetic acid has the sum formula $C_2H_4O_2$ which gives a sum of valencies of $2*4+4*1+2*2 = 16$ which in turn gives 8 covalent bonds. Examining the skeletal structure of acetic acid reveals 8 bonds when the C=O double bond in the carboxylic group is counted twice. We see that it is possible to treat molecular formulae as a formula that summarizes an atomic network.

Senior's aim was to find criteria, which allow for the construction of valid connected graphs out of sets of nodes with a given valency. Connected graphs are networks where each node available i.e. complete molecules. He defined that a graph G is a subset of a partition P, where P is a sorted, non-increasing list of valencies.

He defined:

Z(P)	is the number of distinct graphs for P
ZC(P)	is the number of distinct connected graphs for P
ZL(P)	is the number of distinct loopless graphs for P
ZCL(P)	is the number of distinct connected loopless graphs for P

He stated the following four theorems which are the basis to the statements in Morikawa and Kind.

(i) A necessary and sufficient condition for $Z(P) > 0$ is that $\sum P = 2x$, where x is a positive integer. (remark: x is the number of edges)

If $Z(P) > 0$, then $Ft(P) = \sum P/2-(n-\mathbf{1})$ and $Fr(P) = \sum P/2-p_1$. (p1 is the maximum valency)

(ii) A necessary and sufficient condition for $ZC(P) > 0$ is that $Ft(P) \geq 0$

(iii) A necessary and sufficient condition for $ZL(P) > 0$ is that $Fr(P) \geq 0$

(iv) A necessary and sufficient condition for $ZCL(P) > 0$ is (ii) and (iii)

$\sum P$ is the sum of all valencies and it is twice as much as the number of integers x (atoms) that are associated to the list of valencies. Likewise $n = x$ in the case that P constitutes one connected and loopless graph (loopless refers to the non-existence of edges that connect a node with itself). For a molecule $\sum P = \sum(n_i * v_j)$, which is the sum of all valencies.

Condition (i) is statement 1) in Morikawa and Kind's rule #2. Condition (iii) is statement 2) in Morikawa and Kind's rule #2. Condition (ii) is the condition for the set of nodes to be one connected graph and it is supposed to be statement (iii) in Morikawa and Kind's rule #2. Objectively reading from left to right, this statements says that "The sum of valencies ($2*\sum P/2$) is greater than or equal to twice the number of atoms ($2*n$) minus **1**.", where it clearly has to say "The sum of valencies ($2*\sum P/2$) is greater than or equal to twice the number of atoms ($2*n$) minus **2**.".

While this pitfall in reading Morikawa's statement may appear to be of minor importance, it hinders the following insight that gets apparent on Senior (1951), page 674:

Ft(P) is identical to the cyclomatic number μ if P is one connected graph!

As this is our primary assumption when trying to annotate a formula, it is possible to directly use the cyclomatic number μ for filtering (not μ-0.5 as can be mistaken from Morikawa and

Kind). The cyclomatic number is the number of independently existing circles (paths that lead back to their origin) in a network. Translated into chemistry: μ is the number of double bonds and rings in a molecule.

Why is this knowledge important? It leads us to the first extension of the SENIOR rules: the number of cycles in a neutral molecule can never be negative and it can only be an integer. That means 2μ, the measure which is actually referred to in Morikawa and Kind, can only be even. It changes the nature of the SENIOR rule from being a margin-rule with a continuous positive solution to a rule with a quantized solution!

The following rules result from the cyclomatic number μ:

Let $\mu = \sum(n_i*v_j)/2-n+1$, then if:

(i) If $\mu \geq 0$ and $\mu \in Z$, then the given graph is a neutral, completely connected molecule with μ*(double bonds+rings)

(ii) If $\mu = 0$, then the given graph is an aliphatic molecule.

(iii) If $\mu \geq 0$ and $\mu \neg\in Z$, then the given graph is a completely connected ion.

(iv) If $\mu < 0$ and $|\mu| \neq N$, then the given graph is an aliphatic ion with ammonium functionality; given we investigate neutralized (and neutralizable) molecules, such ions are invalid.

(v) If $\mu < 0$ and $|\mu| = N$, the given molecular formula does not refer to a single molecule and is thus invalid

When annotation of neutral molecules is performed, points (i, iii-v) are exclusion criteria. Chemically, the addition of a cycle is equivalent to the loss of H_2 and therefore μ is inversely proportional to H/C. $\mu = 0$ defines the maximum H/C, however the cyclomatic number cannot differentiate between molecules having a C-backbone (organic) or molecules having a backbone composed of non-C elements. On the other end, there is no definition for a maximum cyclomatic number. Ultimately, large and valid cyclomatic numbers can as well be formed under complete exclusion of H. This is because H has valency $v = 1$, which prevents them from participating in rings. Molecules might also be composed of sterically difficult and invalid structures such as 3-rings. The intramolecular interaction of functionalities may lead to the formation of additional rings, which becomes increasingly problematic when elements have variable valencies. As indicated in table 1, P and S pose a problem. The maximum number of rings that we can use as a filter has to be corrected for functionality interactions; we have to block these interactions. We have to find conventions for:

1) The interaction of functional groups. And therefore the restriction of ‘molecular

backbone formation' with molecules other than C

2) The definition of a maximum number of rings given 1)
3) The definition of a minimum number C's given the cyclomatic number and given 1)

Solution to problem 1)

The sum formulae and the valencies themselves are no basis for a specific restriction of elemental sequences; for this an adjacency matrix for the molecule would be needed (a list of all possible bonds for each elemental pair). Instead, we may abstract the principle of orbital hybridization, which allows sulfur and phosphorus to extend their valencies in order to remove functionalities from the formula. Removing functionalities is the only way to minimize the degrees of freedom for the formation of covalent bonds.

First we assume that P has valency v = 5 and S has valency S = 6. The most common functional groups that support P(5) and S(6) are $R\text{-}PO_4H_2$ and $R\text{-}SO_4H_1$. So if the formula offers a sufficient amount of O and H, which can be dislocated to one of the two functionalities, we can remove either $PO_{(3P+1)}H_P$ (3P+1 for mono-, di-, triphosphate) or SO_4 from the formula. Both functionalities leave one H (v=1) left, which can assume the bond of the former functionality. We can remove as many poly-P or Poly-S as there are O's which can be abstracted into these functionalities.

The study of chemical databases implies, that S(2) may replace O in $R\text{-}PO_4H_2$. For this reason we first translocate all O's and S's into $R\text{-}P(O \text{ or } S)_{(3P+1)}H_P$. Once there is no P left, the other O's are translocated onto SO_4 if possible. Once there are no O's or S's to translocate anymore, we can assume, that the remaining formula has no S(6) or P(5) left anymore. As S(4) plays a minor role and S(2) is more common, we now assume, that any S or P that remains in the formula are either S(2) or P(3). This way, we get a maximum removal of valencies that may interact, while leaving the C-backbone intact. The remaining formula is now assumed to be composed of C(4)H(1)N(3)O(2)S(2)P(3) instead of C(4)H(1)N(3)O(2)S(6or4)P(5), which drastically reduces the number of non-C backbone combinations, given C is existent.

Let us follow the algorithm on the basis of the formula for ATP: $C_{10}H_{16}N_5O_{13}P_3$

The cyclomatic number of $C_{10}H_{16}N_5O_{13}P(5)_3$ is

$\{[(10*4)+(16*1)+(5*3)+(13*2)+(3*5)]-(2*(10+16+5+13+3))+2\}/2 = u = 10.$

Translocating 3P+1 times O and 3P times H onto the number of P's and removing them from the formula results in the formula $C_{10}H_{13}N_5O_3$ whose cyclomatic number is

$\{[(10*4)+(13*1)+(5*3)+(3*2)]-(2*(10+13+5+3))+2\}/2 = u = 7.$

Exactly the three rings associated to triphosphate are removed and the cyclomatic number relates exactly to the number of rings, which relate to ribose = 1 as well as to unsaturations in Adenine = 4, 5-ring in adenine = 1 and 6-ring in adenine =1. That makes 7 rings.

According to Kind and Fiehn, there is no algorithm in existence, which compensates for the multiple valencies of S and P.

Solution to problem 2)

Given our molecule's backbones contain only C and N and given all unnecessary rings were removed, there can be a conjugated π-system which makes roughly n/2 double bonds with n being the number of backbone atoms. In addition, there can be the formation of a six-ring or a 5-ring every six or five atoms. This adds up to $0.5*C+0.2*C+0.5*N+0.2*N = u_{max}$.

Solution to problem 3)

The number of C atoms must be larger than zero and given a backbone of [X-C-X-C-X…], with X being any element of valency $v > 1$, the minimum number of C's cannot be smaller than X-1. (Kirchhoff's rule)

All three remaining rules, the cyclomatic number after the Senior-conditions and elimination of functionalities, the maximum cyclomatic number and the minimum amount of C given a cyclomatic number of zero are to be tested against the seven golden rules in the subsequent section.

Comparison of 7 golden rules versus adapted SENIOR rules.

We have downloaded a set of 18158 exact masses and their molecular formulae from the Pubchem database. We annotated all masses using the in-house written formcalc program, which finds all elemental combinations that satisfy a given error tolerance and a minimum and maximum elemental count. We have used an error tolerance of ± 0.5 ppm. We have applied golden rule #1and allowed elemental counts of 1-70 C, 0-30 O, 0-20 N, 0-10 S and 0-10 P. Formcalc calculated 770067 possible formulae. The application of rule #2 (Lewis and Senior rules as stated in Kind et al) found 250948 acceptable formulae (the correct rule would have yielded 263794 acceptable formulae). Rule #3 (isotope matching) was omitted for above mentioned reasons. Furthermore we applied the semi-strict forms of rules #4 and #5, which pertains to the following elemental ratio filters: $0.1<H/C<6$, $0<O/C<3$, $0<N/C<4$, $0<S/C<3$ and $0<P/C<2$. The application of this filter yielded 222894 formulae. We did not apply the elemental probability filter and since we did not simulate GC-MS we did not apply rule #7.

After application of the relevant filters we obtained 222894 formulae over 18120 true values. In consequence 38 true values (0.21%) of true formulae were omitted.

For the filtering using the adapted Senior rules we have run the following script:

```
% formulae have the form [H C O N S P]
formulae = dlmread('MeSHFormCalc.txt','\t');
N = size(formulae,1);
results = [formulae zeros(N,1)];
valencies = [1 4 2 3 2 3];

for x = 1:N
    formula = formulae(x,:);
    valency = [formulae(x) 0];
    Ores = formula(3);
    Sres = formula(5);
    Pres = formula(6);
    Hres = formula(1);
    Oct = 0;
    SOct = 0;
    Oresct = 0;
    functionality = 0;
    if(formulae(6) > 0)
        for y = 1:formula(6)
        if(formula(3)>=(3*y+1))
            Oct = Oct+1;
        end
        end
        for y = 1:formula(6)
        if((formula(3)+formula(5))>=(3*y+1))
            SOct = SOct+1;
        end
        end
        if(SOct > Oct && Oct >0)
            Sres = (formula(3)+formula(5))-(SOct*3+1);
            Ores = 0;
            Pres = formula(6)-SOct;
            Hres = formula(1)-(SOct);
        end
        if(Oct == SOct && Oct > 0)
            Sres=formula(5);
            Ores = formula(3)-(Oct*3+1);
            Pres = formula(6)-Oct;
            Hres = formula(1)-(Oct);
        end
    end
    if (Sres > 0)
        for y = 1:Sres
        if(Ores >= 4*y)
            Oresct = Oresct +1;
        end
        end
        if (Oresct > 0)
            Ores = Ores-(4*Oresct);
            Sres = Sres-Oresct;
        end
    end
    testform = [Hres formula(2) Ores formula(4) Sres Pres];
    a = testform*valencies';
    b = 2*sum(testform);
    u = (a-b+2);
    umax = 0.5*formula(2)+0.2*formula(2)+0.5*formula(4);
    umax = ceil(umax);
    if(u<0 || u > (2*umax) || mod(u,2)~=0)
       results(x,:)=0;
    end
    if((umax == 0)&& (sum(formula(3:6)==0)))
        results(x,:)=0;
    end
    if ((formula(2)< ((Ores+formula(4)+Sres+Pres)-1)) || (formula(2)==0))
        results(x,:)=0;
    end
end
```

The script first transfers all possible, O and S onto P (to form H_{P+1}(O or $S)_{3P+1}P$. The respective elements are eliminated from the formula. Then all remaining O are transferred onto S so as to give SO_4 and then they are eliminated from the formula. All remaining elements are assumed to have the valency that is stated in line 2 of the script. Based on this formula, the maximum number of rings (umax) and the minimum allowed number of C are calculated.

The application of the script yielded a formula count of 114804 formulae over 17831 true values, which makes a true positive rejection rate of 1.8 %. Relative to the gain in false positive rejection that reduces the final output by 50%, this is an acceptable omission rate.

All the performed annotations were carried out on exact masses with zero error. In this case it would be valid to choose the elemental combination with the smallest absolute deviation from zero ppm. In praxis, it is a common strategy to choose that isobar, which is the closest to zero ppm as being the most likely exact annotation. However, calibration of experimental data can only cancel out the systematic error partition but it cannot cancel out the random error (which then relates to the true accuracy of the instrument). So if the standard deviation of an experimental error distribution is 0.5 ppm, then a measured m/z value can randomly occur anywhere within this field without the error being caused by reason. That means, the choice of an annotation with minimal error has no logical basis for validity of annotation within a given error range. The potential of performing false annotation increases with the number of possible isobaric formulae and it increases by mass.

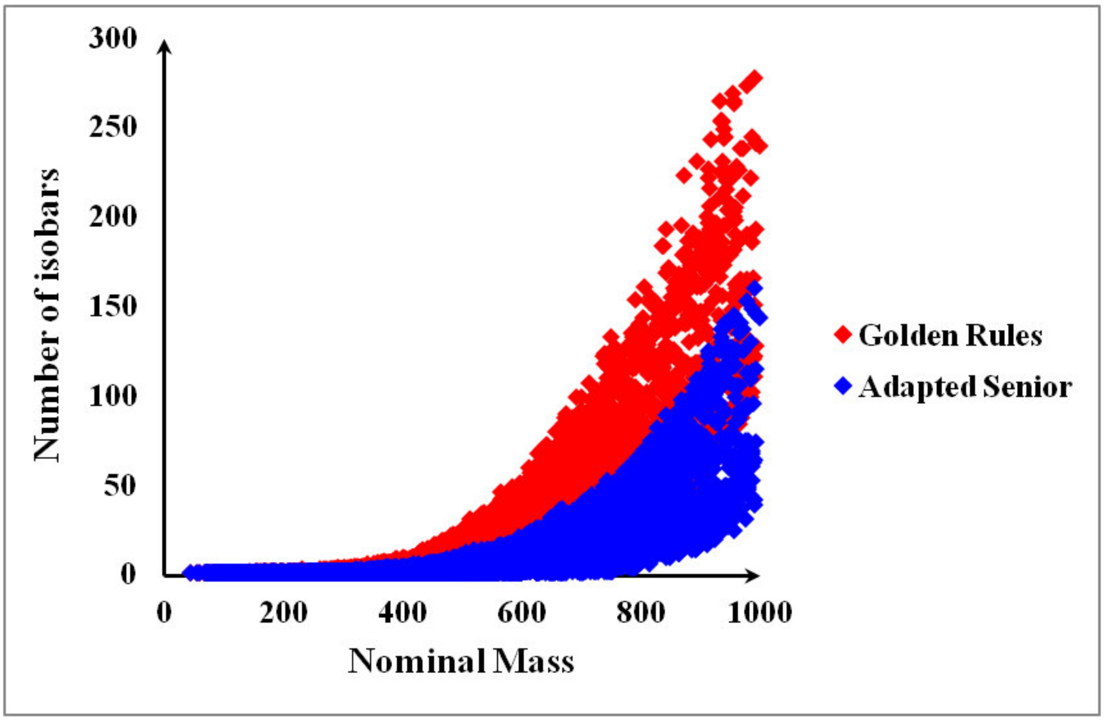


Figure 8: Number of isobars per nominal mass under application of the relevant golden rules and under application of the adapted Senior rules.

According to the above plotted data, rates of presence of isobaric annotations per mass range are as follows:

- 1.3% between 200 ≤ m/z ≤ 300
- 4.6% between 200 ≤ m/z ≤ 300
- 29.5% between 300 ≤ m/z ≤ 400
- 84.1% between 400 ≤ m/z ≤ 500

The probability to make a false decision increases with any new valid isobaric combination and many of them can potentially be true. The same is true for database annotations. Needless to say, the proportion of isobars increases exponentially as error range increases. Unless ultra high accuracy mass spectrometers are used, putative annotation on the basis of mass matching is invalid with a high probability. In consequence, only isotopologue matching can secure correct annotation. As we have deduced above, isotope matching is often no option in non-targeted metabolomics.

Netcalc annotation in connection to the senior rules offers a way out of this demise, because isotopologue checks are replaced by the chemical context of a spectrum. A dense network of non-contradicting stoichiometric relationships supports the probability that a given annotation is in fact appropriate. However, given the large number of alternative isobaric annotation for features of m/z > 400 allows for the coexistence of multiple non-contradicting optima for which reason it is useful to limit the upper m/z margin. Furthermore, an increasing number of

REMDs increases the degrees of freedom for which reason the appropriateness of REMDs has to be evaluated for each dataset.

For the Netcalc annotation of the above dataset we used an EFE of 0.5 ppm and did not limit the final annotation error.

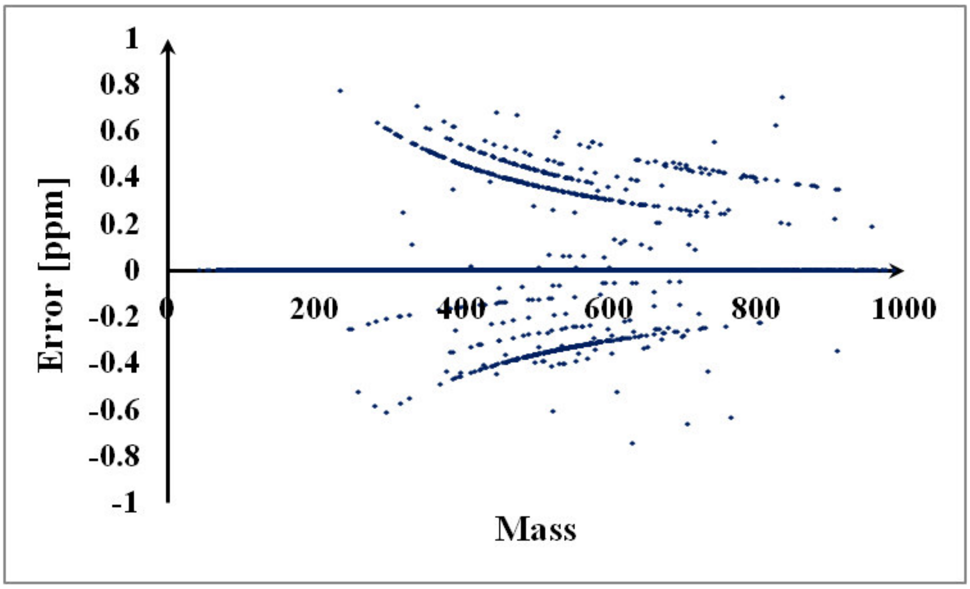


Figure 9: Error-Mass plot Netcalc annotation of the Pubchem dataset.

The Netcalc algorithm yielded 16731 annotations of which 4% were falsely annotated. In consequence there were 15990 of 18158 possible true annotations (88%). Considering an EFE of 0.5 ppm and unlimited annotation error, the amount of false annotations is insignificant. Typical EFEs in Netcalc annotation range from 0.1ppm to 0.2 ppm which usually is equivalent to one to one half error standard deviation. Here, the EFE is infinitely larger than the error standard deviation. Compared to the combinatorial technique there is a significant improvement of annotation quality.

2.5 Metabolic REMDs

Metabolism occurs at many different sites within a cell. Metabolic reactions can take place in cytoplasm or in the plasma of a variety of sequestered cell compartments. These cellular organelles are compartmentalized by membranes which consist of lipid double layers. As metabolism spans all these compartments, metabolites have either to be transported through the lipid double layers in order to meet their next reaction partner or the reaction has to happen associated to the lipid membrane itself. In cases – like the TCA cycle and fatty acid synthesis – reactions are performed by enzyme complexes. Here a substrate is shuffled into the enzyme complex and is subsequently transformed by a series of chemical reactions without the intermediate being released into the cytoplasm at high rates.

Metabolic pathways are a theoretical sequence of reactions, which lead from a compound A to a compound Z and they are formulated to be invariant to cellular structure. No matter how few a chemical analyst might be interested into cellular biology, this cellular biology determines which nodes in a metabolic pathway can actually be detected contemporarily and which not.

Such "omissions of reaction steps" do not only occur on multi-enzyme complexes, they occur on the level of singular enzymes – and in conjunction to coenzymes – as well. An example is the synthesis of Sphinganine: A Serine molecule reacts with a CoA bound residual of palmitic acid to give the precursor of Sphinganine. In this Pyridoxal-PP mediated process a C-C bond between Serine and the palmitic residual are formed [Eliot, A. C & Kirsch, J. F., 2004]. This happens under the elimination of CO_2 and the formal release of H_2O due to its separation from CoA. Translating this reaction into an REMD, one has to consider the original – CoA unbound – substrate of this reaction, the CoA-mediated loss of H_2O and the elimination of CO_2 from Serine. Stoichiometrically it is a two step reaction. So even if all amino acids and all fatty acids were considered given an ordinary condensation, without the definition of the proper reaction type, it would be impossible to capture this described reaction. In the present thesis this reaction is denoted as "decarboxylative condensation". Especially in conjunction to Pyridoxal-PP (vitamin B6) such multi-step reactions occur.

Other reactions which are mediated by Pyridoxal-PP are forming C-C bonds among alpha-C atoms under de-carboxylation. In this thesis this rather exotic reaction is denoted as "decarboxylative addition". Independently of CoA it may occur between α-keto acids and e.g. amino acids.

Other reactions that can occur in conjunction are hydrogenation and condensation. Glutamate and a number of Oxo-acids are involved into such reactions, which given their free substrates can be formalized as A + B = C + Oxygen. Here, this reaction is called "condensation on hydrogenated carbonyls".

Interestingly, Pyridoxal-PP is involved in all reactions which are not regular condensations/hydrolyses or oxidations/reductions involving O, PO_3 or SO_3. Paradoxically, in literature Pyridoxal-PP has a negligible role as compared to CoA.

Pyridoxal-PP, together with NADPH, plays a major role in deamination and transamination. Due to the change in ESI-ionizability for potential substrates and products of these reactions, all reactions which pertain to amino acids (majorily condensations) were as well formulated keeping their transamination products in mind.

Needless to say, the discontinuity that fatty acid synthesis might evoke, was compensated by listing all reaction types, condensation, ‘condensation on hydrogenized carbonyls’, ‘decarboxylative condensation’ and ‘decarboxylative addition’ onto any intermediate product of the fatty acid synthesis cycle.

The same was done for amino acids and their respective keto acids. Another class of reactions considered encompasses conjunctions with common metabolites or lipid head groups.

In its entirety, all considered reactions (REMDs) amount to a number of 175 reactions. The complete set of REMDs is listed in the supplementary material.

Analytical REMD domains:

The set of REMDs described above mimics metabolism and eventual stoichiometric gaps. Taking ESI-ionizability of different metabolites into account, all above mentioned reactions can only be detected within the same domain of ions.

Netcalc only annotates m/z peaks as a function of the specificities to the starting m/z and the REMDs. That means that it does not annotate isotopologues and can conversely be used as a de-isotoping tool which ultimately reduces data co-linearity. But inherent to this property, Netcalc cannot annotate $[M+Na^+]^+$ ions if annotation was started from $[M+H^+]^+$ ions unless the mass which connects both domains was specified.

In ESI ionization $[M+H^+]^+$ ions and $[M+Na^+]^+$ ions dominate the positive ionization mode. $[M-H^+]^-$ and if Cl is present $[M+Cl^-]^-$ ions dominate the negative ionization mode. That means, in order to capture every annotatable metabolite in a sample the following transformations have to be added to the metabolomics list:

- The transfer from H^+ to Na^+ ions with no change in CHNOSP
- The transfer from $-H^+$ to Cl^- ions

If annotation outside of this domain of 6 elements and optionally two synthetic elements ($|H^+-Na^+|$ and $|H^++Cl^-|$) is desired, all new “elements” such as ^{13}C, ^{2}H, ^{15}N, ^{18}O, ^{34}S or ^{32}P need to be specified as being new elements. In addition each new element has to be attributed with a filter in reference to the other elements.

In conclusion, the metabolomics REMD list needs to respect both, possible analytical gaps in respect to metabolic pathways and analytical gaps in respect to ion types. Interestingly, Netcalc also differentiates molecular adducts, such as MeOH or NH_4^+ adducts, as long as the change of domain is not specified. The complete list of metabolomic REMDs can be found under supplemental information.

2.6 Analysis of Error Sources

2.6.1 Netcalc versus Database Matching

All main annotation strategies – combinatorial, database matching and Netcalc– have their inherent advantages and disadvantages. It is important to know all these techniques' properties in order to judge, whether they are used and interpreted in an appropriate manner. However, they all have things in common. All techniques are bound to a solution-search within an error window, which is assumed to be befitting to the analytical instrument and the analytical procedure, and the plausibility of the uniqueness of a given finding depends on the respective instrument's resolving power.

The combinatorial technique

The combinatorial technique produces solutions of elemental combinations which as close as possible match the experimental mass. We have shown above, that the list of solutions for different masses is vast and encompasses a large amount of isobars (elemental combinations which fit to the same mass in a given tolerance window). The choice of the correct isobar depends on the detection of isotopologue peaks whose intensity has to match the expected magnitude of a given molecular formula. By experience, however, the fewest users of mass spectrometry are aware of the fact, that mass defects of isotopologue peaks, whose difference to their most abundant isotopologues is nominally identical differ in mass defect; e.g. nominally: ^{13}C-^{12}C = 1 Da and ^{15}N-^{14}N = 1 Da but exactly: ^{13}C-^{12}C = 1.003355 Da and ^{15}N-^{14}N = 0.997035 Da. At m/z = 400 a resolving power of R = 63 291 is needed to resolve this case mass spectrometrically. There are various combinations of isotopological isobars which cannot be resolved at common resolving powers of R<< 100 000. The worse mass spectrometric accuracy, the more combinations which may fit to a given mass can be found and the lower the resolution, the worse they can be distinguished. In addition, given that mass accuracy and resolving power are sufficient, the mass to be annotated needs to be present at an abundance sufficiently high to detect the isotopologues because at m/z < 1 000 and z = 1 isotopologue peaks can be less abundant than the exact mass peak by two orders of magnitude (and even more). This is problematic in regard to the power law distribution of m/z peak intensities. Consequently the molecular formula of only a small amount (< 5%) of m/z peaks can be verified at hand of their isotopologue abundances.

The database technique

The advantage of this technique is that it allows for direct access to the biochemistry, pathway affiliation and literature relayed for a given hit. However, the drawback is, that database entries – even if there is a hit in the pre-specified error window against which the query was run – may result in a 'best guess in default of a better solution', since databases do not cover the entire combinatorial space of possible annotations. The effect of this 'best guess in default' mechanism can easily be depicted by observing the error distributions of annotation results at different ppm. In chapter five, we perform a Netcalc annotation which results in several thousands of hits. We translate the experimental masses into exact masses and perform database matching using MassTRIX [Suhre. K. & Schmitt-Kopplin. Ph, 2008] at 0.1 ppm.

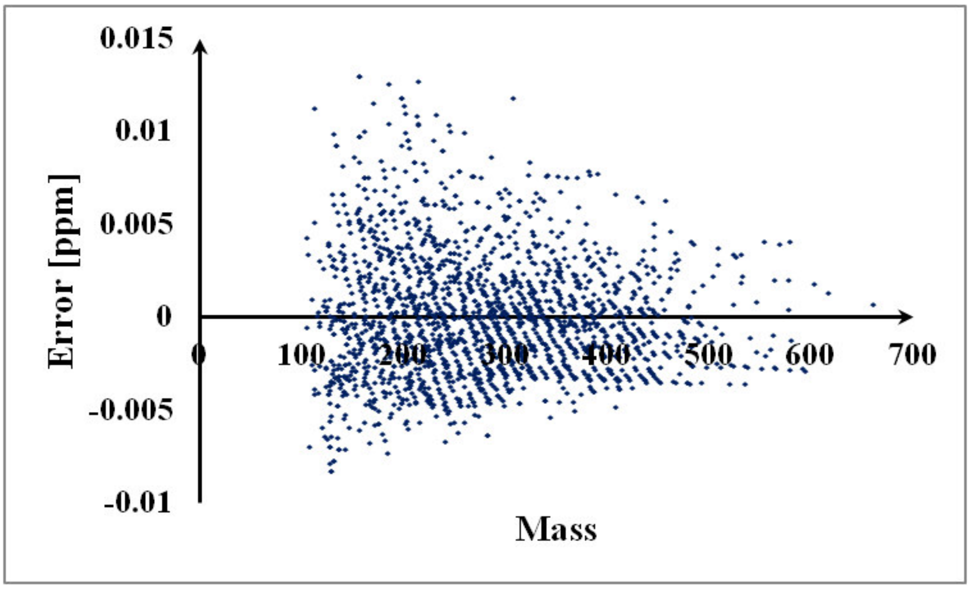


Figure 10: Error over mass plot of MassTRIX annotation (0.1 ppm) of theoretical masses derived from previous Netcalc annotation.

The resulting 2006 annotations are majorly not spreading beyond 0.01 ppm, which is 10% of the assigned tolerance. The given annotations differ by rounding errors and we have 100% of formula matching between Netcalc and MassTRIX. Increasing the error margin results in a change of this scenario.

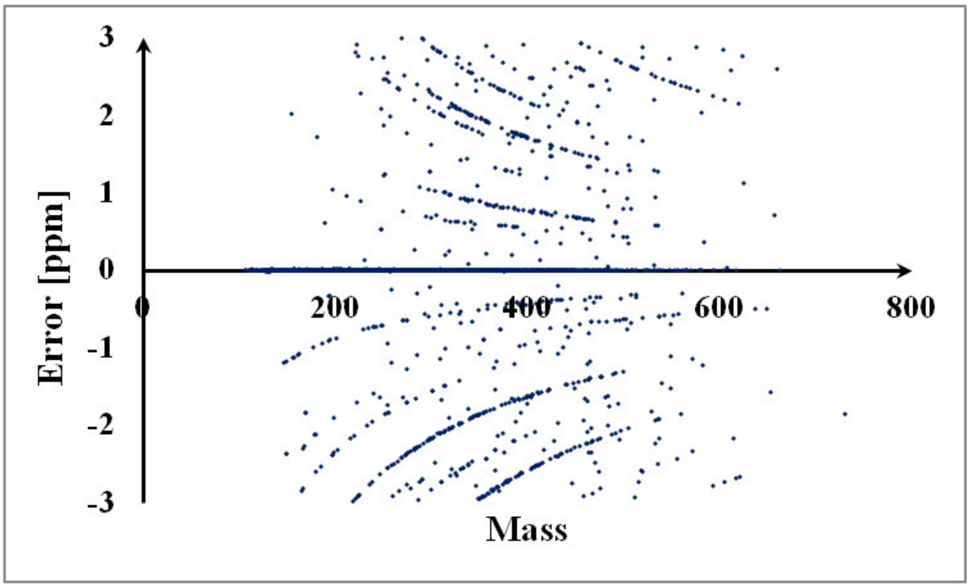


Figure 11: Error over mass plot of MassTRIX annotation (3 ppm) of theoretical masses derived from previous Netcalc annotation.

MassTRIX annotation at 3 ppm error tolerance shows an annotation of isobars throughout all ppm-ranges and we found a proportion of false annotations of 27.6%. The picture looks very structured because the annotation was performed on theoretical values. Nonetheless, we can observe, that the next closest annotation is assigned to the theoretical masses because the respective true value is not listed in the database. Applied on experimental datasets the same test results in a fuzzier picture because the experimental masses are spread along the y axis. Compared to the Netcalc annotation of the Pubchem dataset (infinite error tolerance and EFE = 0.5 ppm) the proportion of false annotation is seven times larger in the MassTRIX annotation and the absolute number of actual true annotations is eight times smaller.

In consequence, we can state, that the optimal strategy for the analysis of metabolomics datasets is to first perform Netcalc annotation and to then perform database matching on the yielded theoretical masses.

The Netcalc technique

The Netcalc algorithm has the following advantages as compared to the combinatorial approach and the database matching approach:

- The combinatorial space is completely covered, just as in the combinatorial approach, but an m/z peak annotation is only then accepted, when all its stoichiometric relations to all its adjacent m/z peaks are consistent with the REMDs that connect them. One may call it “democratic adjustment” which makes the formula valid
- It does not need isotopologue peaks since “the community” makes the decision as to

which an isobar is the correct annotation

- Netcalc annotations happen in small steps throughout the m/z range, in error steps that are much smaller than the tolerance windows specified for the other two approaches. Depending on this step-size and the density of annotatable peaks, Netcalc annotations cannot move far away from the original error distribution and therefore give a more realistic guess upon an m/z peak's formula
- The network character of Netcalc gives rise to a new foundation of pathway analysis without being restricted to reference pathway maps. Given that the metabolic REMDs cover all possible metabolic reactions, metabolic pathways are guaranteed to be a subset of the Netcalc-network for as long as the reactands are detectable

Nonetheless, Netcalc is vulnerable to false annotations as well. Naturally, an increasing number of elements that are used for the calculation of a molecular formula increases the number of isobaric annotations. In theory, one property of Netcalc is that it cannot annotate an m/z value that has no REMD associated to it. In consequence, if annotation is started with a protonated mass $[M+H^+]^+$ and the transition to sodium adducts (Na^+-H^+) is not defined, Netcalc cannot annotate ions of the form $[M+Na^+]^+$. Likewise, if transitions to isotopologues are not defined and we start annotation with the exact mass formula, no isotopologue can be annotated. However, we have observed that $[M+Na^+]^+$ molecules can acquire annotations despite the lack of a $[Na^+$-$H^+]$-REMD. For example: Glucose $[C_6H_{12}O_6+Na^+]^+$ can attain the annotation $[C_{12}H_{10}OS+H^+]^+$ even if $[Na^+$-$H^+]$ is not defined. The difference of both annotations is 0.48 ppm. The respective 'mis-annotation' was observed to take place even at an EFE of 0.1 ppm.

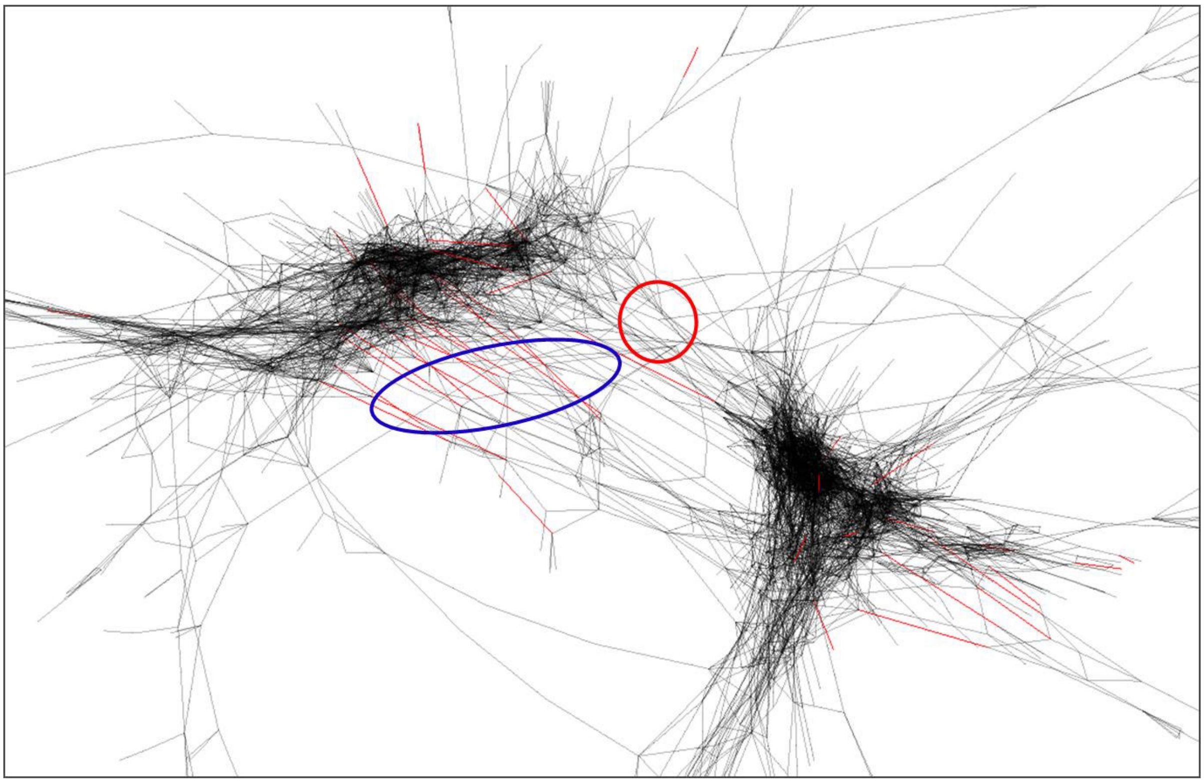


Figure 12: Mass difference network. Two modules are visible. One module pertains to $[M+H^+]^+$ ions, the other module pertains to $[M+Na^+]^+$. Red edges highlighted by the blue ellipse are $[Na^+- H^+]$ REMDs, which allow the transition from proton space to sodium space. Black edges highlighted by the red circle are invalid connections between both ion spaces.

Figure 12 shows that connections between ion-spaces exist even though the initial assumption underlying Netcalc is that an REMD uniquely addresses one stoichiometric relationship.

The original conclusion that was drawn from this phenomenon was that non-allowed paths between ion-spaces exist because of insufficient mass spectral quality and insufficient calibration, which lead to false edge assignments. It was assumed that singular peaks – especially close to the noise – have random mass shifts which are strong enough to move into the position of another theoretically existing peak.

This hypothesis was supported by an experiment in which we annotated the same mass spectrum at different EFE levels. Then we extracted the respective sparse matrices, and counted true positive assignments (TP) and false positive assignments (FP). TPs related to mass pairs, whose formula difference matched their incident REMD. FPs related to mass pairs whose formula difference did not match their incident REMD.

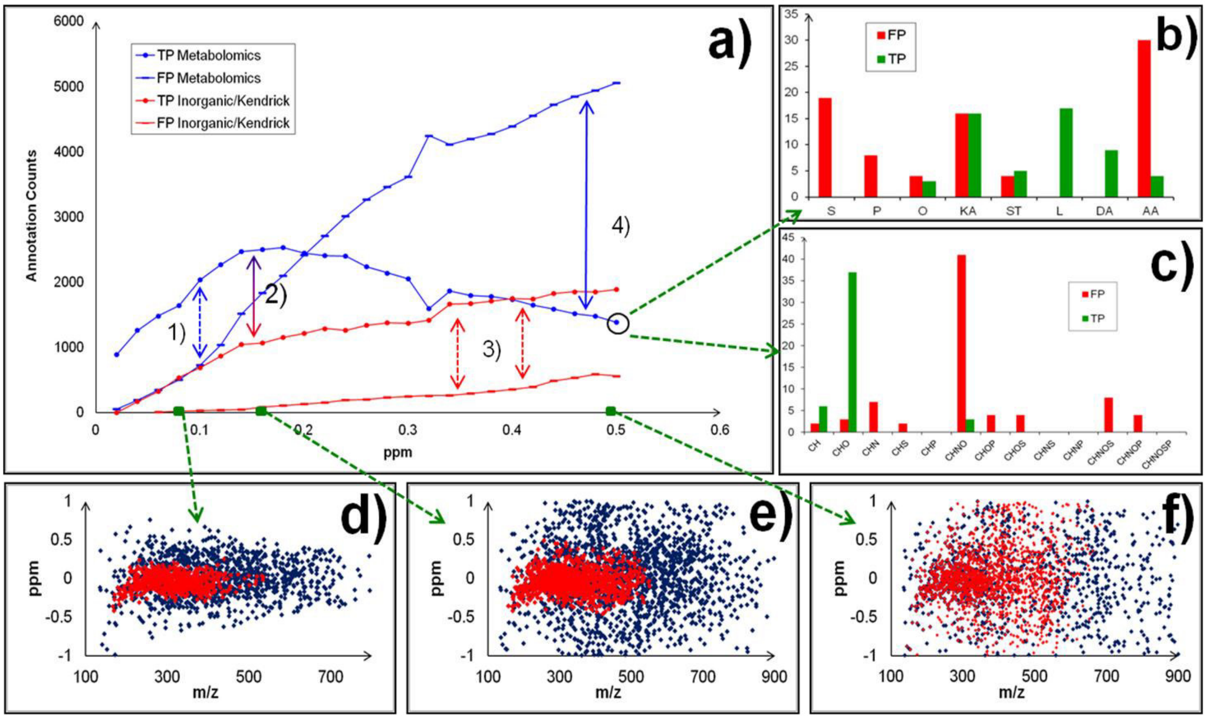


Figure 13: a) Counts of masses involved with false positive mass difference assignments (FP) and true positive mass difference (TP) assignments; a1) Greatest distance between TP and FP in metabolomics list; a2) Greatest distance of TP assignments between metabolomics and structural list; a3) TP to FP distance is increasing for the structural list as a function of networking error; a4) Increasing divergeance between FPs and TPs as a function of networking error in the metabolomics list. b) Metabolomics list| 0.5 ppm: Abundance of TP and FP mass difference assignments among reference mass difference groups [S = Sulphur Organic; P = Phosphorus Organic; O = Other Organics; KA = Keto Acids; ST = Structural; L = Fatty Acids and Isoprene units; DA = Dicarboxylic Acids; AA = Amino Acids]. c) Elemental compositions as a function of FPs and TPs. d) Red = Structural| Blue = Metabolomics| error distributions at 0.08 ppm networking error e) Red = Structural| Blue = Metabolomics| error distributions at 0.16 ppm networking error. f) Red = Structural| Blue = Metabolomics| error distributions at 0.5 ppm networking error

In addition we had grouped REMDs into different classes in order to find out, whether specific REMD classes were especially prone to FPs. We found that mass differences containing sulfur (S), phosphorus (P) or amino acids (AA) were almost exclusively involved into false positive edges. Lipids and dicarboxylic acid REMDs in turn were especially pronounced throughout TPs. In addition, figures 13d, e and f show that the error distributions yielded from annotations with the Kendrick REMD list (published by Tziotis in 2011) and the metabolomics REMDs (introduced in chapter 2.5) dissociated at different error levels. The metabolomics REMD list appeared to be a major source of false annotation at EFEs larger than 0.1 ppm.

The results gained from this experiment supported the assumption that random events as well as sulfur and phosphorus containing REMDs were causative for false edge formation.

Further investigations upon the given problem revealed, a mistake (or an inaccuracy) in the filtering function. Given an absolute error tolerance of 1 ppm, the initial filtering function said ‘If $\Delta|m_{experimental}$, $[M+H^+]^+|$ < 1ppm or if $\Delta|m_{experimental}$, $[M+Na^+]^+|$ < 1ppm, then accept the formula. This line allowed the algorithm to choose between two possible solutions and therefore increased the degrees of freedom of the filtering process. Realizing this mistake, we explicitly specified Na as an element for formula calculation. As a consequence the algorithm only had one option (false or true) which improved annotation performance tremendously.

Mimicry Mass Moieties

The final hint as to why non-allowed edge assignments appear, was found by analyzing the maximum distances between theoretical masses of the same data set that was used for chapter 2.4. We produced the respective $[M+H^+]^+$ masses and $[M+Na^+]^+$ masses, concatenated them and sorted them in increasing order. We found that there is a conserved set of elemental combinations that – if added onto an REMD – yielded formula differences whose masses are so similar to the actual REMD that they cannot be differentiated from the original REMD unless 0.05 ppm accuracy is achieved.

In other words, for almost any REMD there is a naturally occurring equivalent formula of positive and negative elemental counts, that sums up to the same mass within an error range of << 1 ppm (and even < 0.1 ppm). We can refer to such elemental compositions as 'mimicry moieties'.

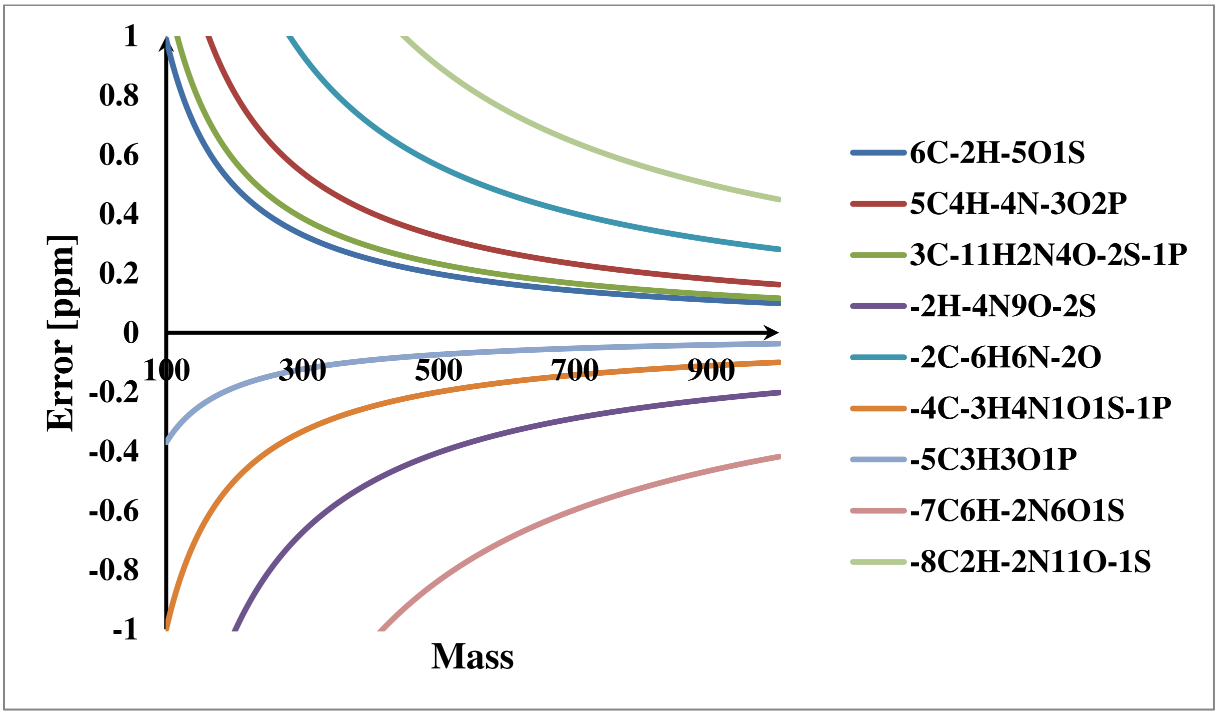


Figure 14: 10 different elemental combinations which almost exactly match the mass of $[Na^+-H^+]$ – the transition from proton space to sodium space. The combination '-5C3H3O1P' is closer than 0.4 ppm to $[Na^+-H^+]$ throughout the entire mass range. In addition, this mimicry moiety induces minimal changes in the isotopic pattern of higher masses because the isotope ^{13}C is not abundant enough to induce notable changes in the isotope pattern.

Aside of the above displayed sodium-like moieties, there are zero mass moieties, whose combination simply adds up to a number that is close to zero. The term ‘zero mass moieties’ was introduced by M. Perdue in an oral presentation at Helmholtz Zentrum München in 2010), Many mimicry moieties are simply the sum of a valid REMD and a zero mass moiety.

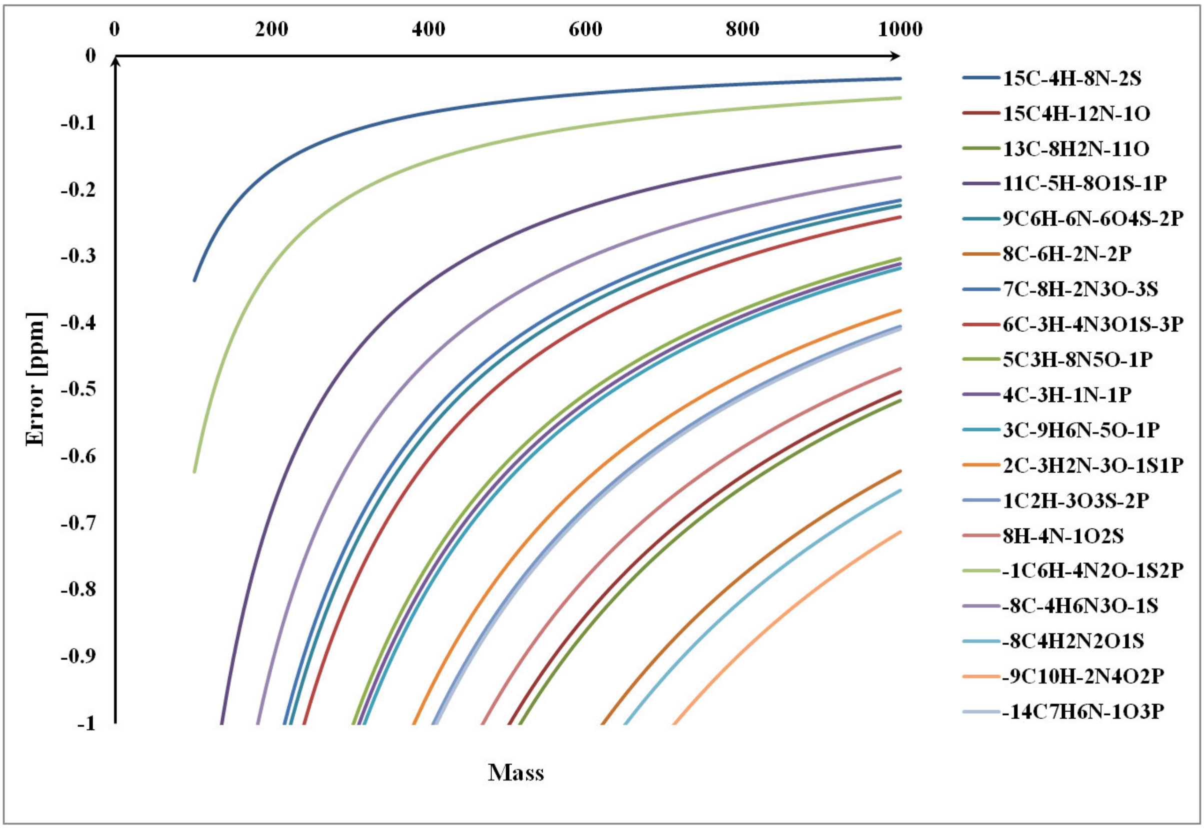


Figure 15: 19 elemental combinations, which almost exactly match the zero mass.

Any REMD can potentially be substituted by a mimicry moiety. A closer investigation of the mimicry moieties reveals that almost every mimicry moiety has proportions of P and S. The reason for this phenomenon lies in the mass defect of each element. H and N have positive mass defects (0.007825 and 003074, respectively), while O, S and P have negative mass defects (-0.005085, -0.027929 and -0.026238, respectively). A certain number of H and N it is necessary to add a sufficient number of O, S or P in order to balance the sum of masses around zero. For this reason REMDs which are associated with S or P are typical for ‘false positive’ edge assignments. In consequence, any Netcalc approach whose REMDs are composed of C, H and N exclusively, cannot contain such mimicry moieties.

Is it true then, that falsely assigned REMDs indicate false annotation? This is in fact not necessarily the case, because we have extracted the mimicry moieties from a mass list of valid formulae. Netcalc annotation works by walking along the edges within the mass difference network. Two network branches may be correctly annotated but they may contain formulae, whose mass difference is close to that of an actual REMD. That means the formulae may be correct, even though another REMD is assigned to them. However, unless we have the possibility of isotopologue matching, we cannot prove either of the given scenarios.

For that reason, we have decided to randomly permute the search directions through the network and to count how often an REMD was assigned inappropriately. An inappropriate assignment of an REMD causes the deletion of its incident annotations and the algorithm has another chance to find a less contradicting solution in the next iteration. However, since multiple solutions with minimum contradiction may exist, the algorithm may oscillate ad infinitum. For this reason we have decided to list the average annotations of several read outs once oscillation has begun, because this way it is possible to devise a set of most likely correct annotations.

3 Data Analysis 1: Development of appropriate Workflows for Data Cleaning

3.1 Data Handling

The search for surrogate markers in metabolomics data is impeded by the complexity of such data and by the claim, that such markers are expected to have equal and preferentially better performance than known markers. A multitude of challenges arises from this claim, because the regime that governs uni-variate and multi-variate statistics is theory-oriented rather than empiricism-oriented. In consequence, the majority of approaches try to bend empirical data in order to fit theoretical standards and if this is not possible, data is discarded. Metabolomics data has properties, which conflict with statistical standards:

1) the excessive number of missing values, which are replaced by zeros
2) Occurrence of multi-modal data
3) Non-normal intensity distributions
4) Occurrence of co-linearity

As a consequence of these factors, typical normalization techniques – parametric or non-parametric – as well as datamining techniques, such as PCA and correlation networks, fail to correctly extract and interpret the given data. Therefore, datasets are either discarded – urging the experimentalist to repeat his measurements – or more complicated techniques such as variations of PLS, random forest clustering, SVMs or SOMs are applied. The drawback of the latter is that they might indeed lead to a good data separation, but the cause of this separation often is unclear.

The lack of understanding as to ‘why’ and ‘how’ a variable is important for the differentiation of data leads to the impossibility to show this separation in the original data, which ultimately induces distrust that a scientist has towards his own results [experience shared with a wide circle of colleagues]. In addition, the unawareness of the nature or mechanism as to how a variable is significant can induce false interpretations of data.

3.2 The Matrix Effect

The analysis of any biological matrix in direct infusion MS is complicated by common effects, but each matrix in particular poses specific challenges in addition. In targeted analysis, the effect which can confound a target peak is called matrix effect. This term describes the interfering action of all other constituents of the analytical matrix, which can be controlled by developing a method that removes most compound of non-interest. Further sources of complication can be the composition of solvents, dead volumes in the analytical system, which may lead to the accumulation of matrix or errors in sample cleanup, which is dependent on the experimentalist.

In non-targeted analytics, however, we analyze the matrix in its entirety; the matrix effect is therefore an inevitable phenomenon in metabolomics measurements. Since we cannot control the matrix effect by the deletion of “the matrix”, we have to develop data analytical strategies, which allow for the integration and interpretation of matrix effect. In fact it is important to exactly define which partition of the data relates to the biological matrix and which partition relates to impurities, artifacts or contaminations.

In contrast to other biofluids such as plasma, CSF and urine as well, EBC is characterized by a very strong variability of “matrix content” per sample. The matrix – the metabolome – in EBC is carried by aerosolic droplets, which derive from the airway lining fluid. Exhaled H_2O (g) is co-condensed during the sampling process and serves as a solvent for the droplets. The amount of exhaled H_2O (g) varies strongly as a function of body-hydration and metabolic activity. Other than urine, EBC stems from an entirely open system and its analyte concentrations are reported to vary more than 5-fold intra-specifically and inter-specifically.

How varying analyte setup and varying sample dilution – a multi-parametric setup – affects the quantitative output of ionization sources and how this output can be handled data analytically, is poorly documented. However, this knowledge is indispensible for the production and interpretation of results. The traditional mindset of bioanalysts is very much focused on univariate measures, such as standard deviation, relative standard deviation (RSD), significance testing, confidence intervals and many more. Obviously, the multi-parametric nature of EBC, and matrix interactions render such measures to be inappropriate unless these parameters can be controlled.

3.3 Specifying the Problem

3.3.1 Varying Feature Counts

How do matrix effects and dilution influence traditional data correction? Let us assume that the metabolome in a sample type counts 2000 metabolites and that an optimal analytical setup will reveal exactly 2000 m/z signals. Let us then assume that another sample of the same sample type is diluted in a manner that leads to a detection of the most frequent 1000 metabolites only.

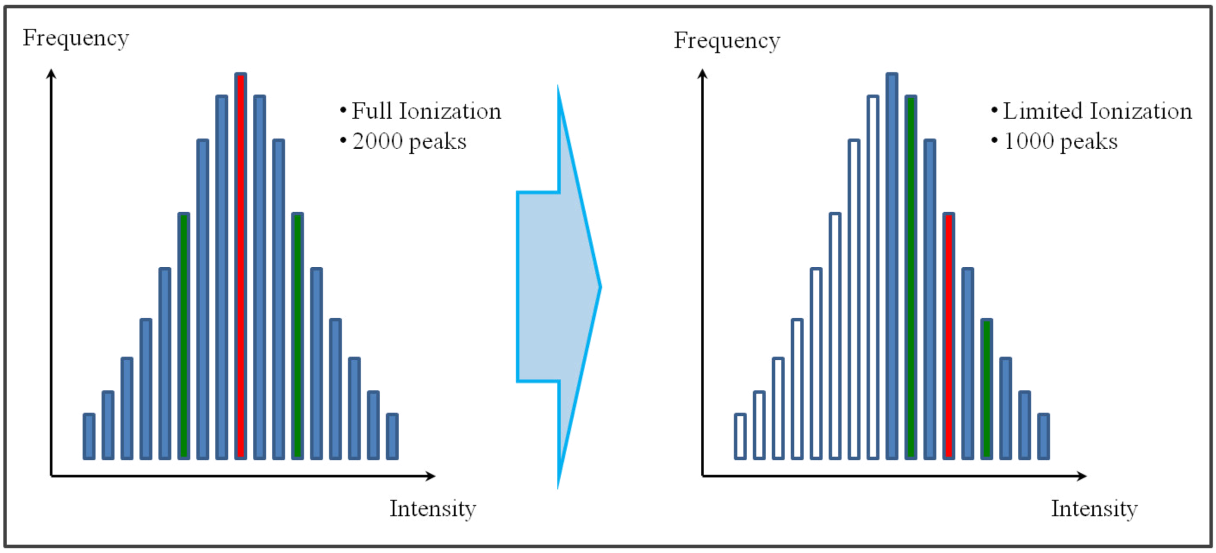


Figure 16: Hypothetical Intensity-Frequency histogram referring to full ionization of 2000 metabolites (left) and Intensity-Frequency histogram referring to limited ionization yielding 1000 metabolites. The relative order of intensities in the samples is assumed to be constant. The red bar marks the median, green bars mark the first and the third quartile, empty bars relate to signals that get lost due to limited ionization.

Let us investigate the effect of different normalizations based on this scenario. If the data is normalized on the inter quartile range, then the 3rd quartile in the full ionization scenario will become the median in the limited scenario. It is obvious, that metabolites which belong to the same intensity bar will get over-estimated in the limited case. The only metabolite that will not be affected is the maximum abundance metabolite.

The scenario for non-robust normalizations such as Manhattan norm, Euclidean norm, Maximum norm, normalizations on standard deviation or variance all lead to the same result given the above scenario. A technique that is often applied in order to make the intensities of features comparable is called scaling – in fact normalization of the independent variables. However, different feature counts cause non-random missingness of feature signals throughout the samples. Scaling does not solve the problem, because the different feature counts per sample stay. In consequence, there is a dilemma: Whether we normalize first or whether we scale first, does not solve the problem.

3.3.2 Power Relationships and Scale-free Relationships

A further problem which pertains to omics data – and to non-targeted metabolomics data in particular – is the power distribution or scale-free distribution of metabolite ion abundances, which is also enhanced by differential ionizabilities of different compound classes.

Power relationships describe a scenario, where one measure varies by a power with the variation of another measure.

The power law can be written as: $f(x) = ax^{k}$.
Scale-freeness has negative values of k: $f(x) = ax^{-k}$.
If f(x) is the probability of occurrence: $P(x) = ax^{-k}$.

The scale-free relationship in words; if we assign the intensities of a mass spectrum into intensity bins, the probability to find a peak, which we can assign to the maximum intensity bin is by the power of –k smaller than the probability to find peaks that can be assigned to the second largest bin [Willinger, W., et al. 2004; Li, L., et al. 2005].

The power relationship in words: The intensity of the maximum peak is k times larger than the second largest peak; the second largest peak is k times larger than the third largest peak and so on.

Intensity distributions in metabolomics data often follow such a relationship. This relationship, however, poses large problems for data pre-treatment, because if a norm is formulated over non-scaled data, very few peaks can impose the majority of weight onto the norm. The scenario becomes intriguing if one of the features with largest intensity is actually a univariate marker. Normalization would then impose the inverse pattern on all other features; normalization would then ‘fake’ statistical significance. A norm must therefore always be tested for independency between sample groups of interest.

The dilemma deriving from the missingness structure is magnified here, because powerdistributions imply scaling prior to normalization.

The problem faced here, is very specific to non-targeted metabolomics; genomics data and targeted metabolomics data are commonly performed using full set analysis, where such problems do not occur. For this reason present literature contains almost no workflows, which treat this problem.

3.4 Analysis of Dilution Series

In this section we analyze a dataset, which was specifically composed to address the following characteristics:

- varying dilution
- sample pre-treatment

The given dataset consists of technical triplicates of a six-point dilution series of one EBC sample, a technical triplicate of another EBC sample and two point dilution series of divers SPE treatments. The aim of this experiment is to test existing data analytical strategies for their capability to minimize the difference of metabolite levels throughout the triplicates along the six-point dilution series. We chose to optimize this parameter, because all these 18 samples are essentially different versions of one and the same sample. A potent data treatment strategy should enable the researcher to identify these samples as being one and the same; both, in a uni-variate manner and in a multi-variate manner.

3.4.1 Analysis of the six-point dilution series

At first, we analyze the triplicates of the six-point dilution series and a methanol blank (infinite dilution) in order to investigate which different responses analytes can have towards dilution. We apply different commonly used data-treatment methods in order to investigate, which effect they have on the data, and whether it is possible to cancel out the dilution effect solely by the use of this set.

Prior to the application of any techniques of datamining or statistics, it is useful to get an overview over the data. Figure 17 shows a mass defect over m/z plot. Based on the element specific nature and number of covalent bonds exact masses of CHONSP metabolites cannot exhibit arbitrary mass defects (the digits after the comma). In consequence there are forbidden zones whose population would indicate the presence of non-CHONSP molecules.

Figure 17 demonstrates the absolute mass defect (AMD) distribution in dependency of compound mass (or m/z) of 18 159 manually downloaded compounds in comparison to the dilution series data. The population of the forbidden zone can usually be attributed to doubly charged ions, salts or random noise. The present data was mined with S/N = 4. Lower S/N thresholds usually lead to a population of the forbidden zone. We can see, that the dilution data series mostly populates higher AMD allowed regions, which indicates low oxygenation, low amounts of salts or chloride adducts. Molecules can be predicted to be mostly consisting of CH and N, whereas they need to have at least one O or S in order to be ionizable in negative mode.

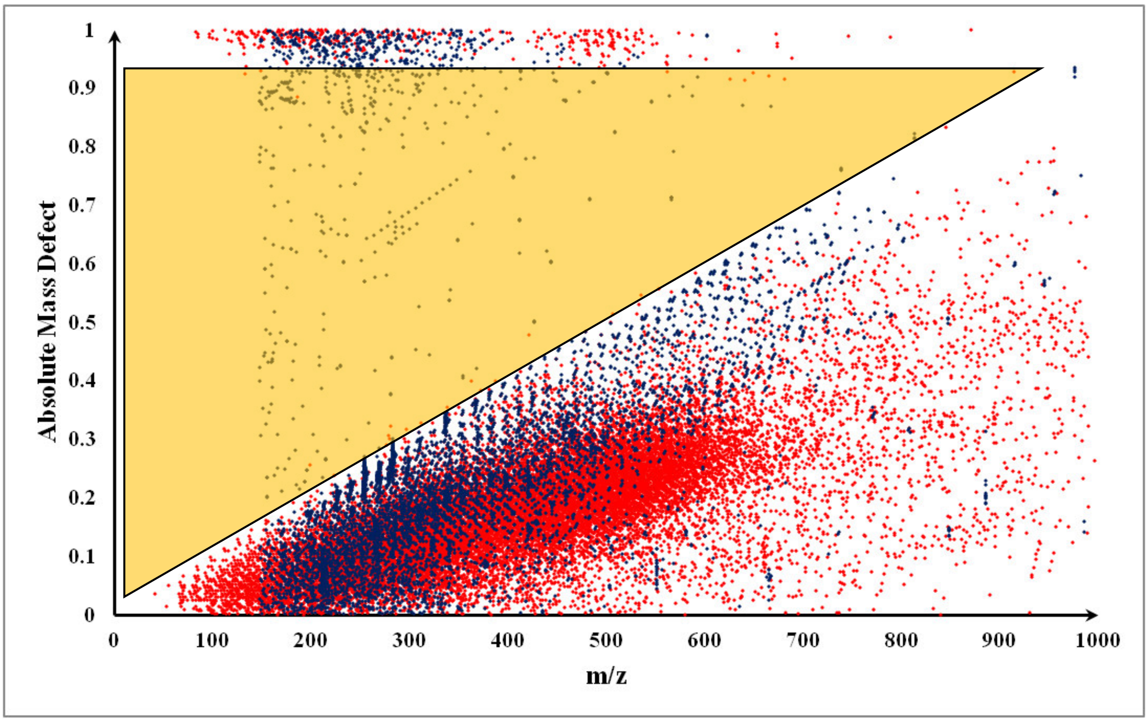


Figure 17: Absolute Mass Defect over m/z plot. Red dots are 18 159 CHONSP compositions downloaded from Pubchem, corrected for proton removal; Blue dots are the EBC dilution series data. The orange Triangle indicates the forbidden zone.

The next test on the raw spectra is used in order to investigate the relationship of neighboring m/z peaks. The Fourier transform, which generates the mass spectrum from the frequency domain, causes large peaks to show “wrinkles”; the peak “oscillates in” and “oscillates out”. These wrinkles are often larger than the given S/N margin and they are usually direct neighbors of another. Wrinkles are problematic because they may be falsely annotated, thereby leading to false interpretations and – even more importantly – they introduce co-linearity into the dataset. In multivariate analysis, excess colinearity may polarize the correlation structure within the data, thereby “suppressing” information of interest. In univariate analysis, using multiple testing they may constitute a set of markers without adding relevant information to the marker set. Wrinkles can be extracted by sorting the mass-sample-intensity matrix for m/z and by correlating any i^{th} entry with its corresponding $i_{+1}{}^{th}$ entry.

Wrinkles (and double peaks) will exhibit correlation coefficients close to 1.

Figure 18 shows a neighbor-correlation versus m/z plot. Inspecting the region of $r^2 \approx 1$ we can see, that the amount of wrinkles and double peaks is fairly small. Overall we can verify good mass spectral quality. Strong miss-alignment of mass spectra (poor calibration) can be deduced from Pearson correlation coefficient close to -1. This is because a miss-calibrated mass shifts into the next adjacent mass-line which yields an alternating zero-non-zero pattern over the samples.

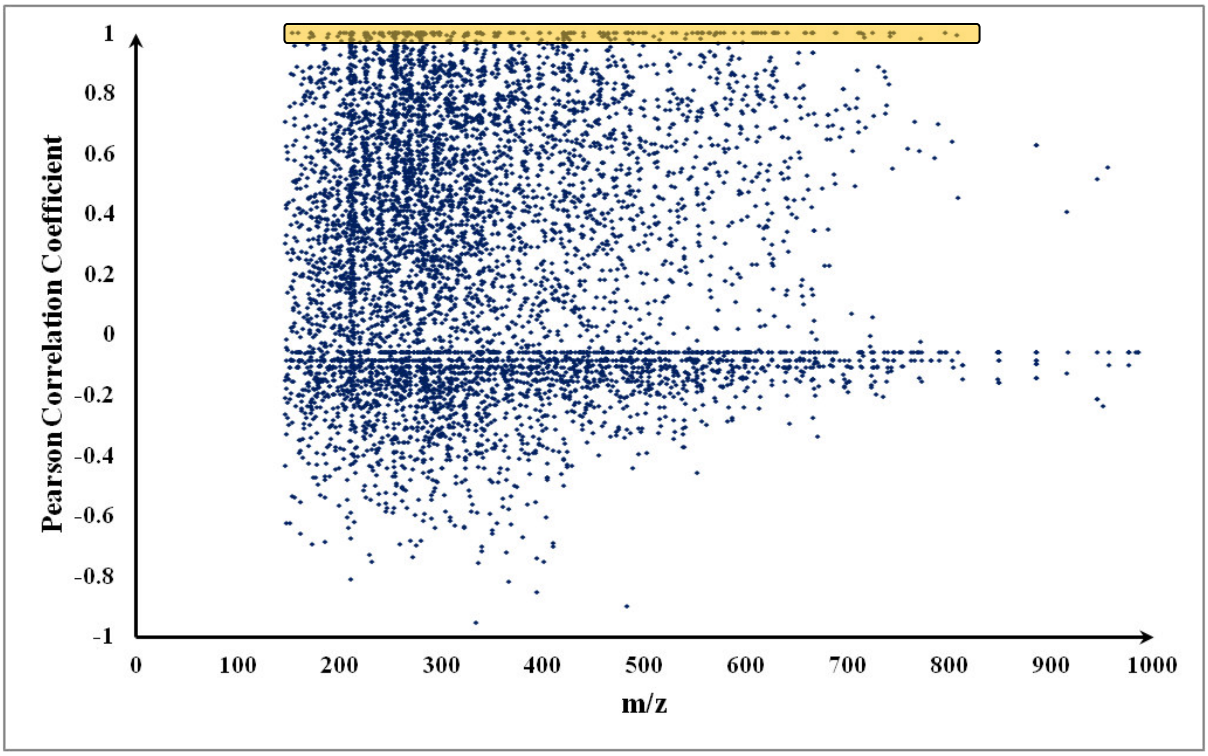


Figure 18: Running Pearson correlation coefficient over m/z. The yellow rectangle shows peaks with wrinkles or peaks splitting.

The running Pearson correlation plot of our dataset indicates acceptable mass spectral quality in terms of calibration and colinearity; the mass-sample-intensity matrix can now be analyzed for its response to dilution.

The initial “naïve” prediction of the response of the analytical matrix towards dilution would be that both, peak intensities and peak counts decrease with increasing dilution factor. We test this assumption by plotting values representative for the overall intensity distribution of each mass spectrum over natural logarithm of their respective dilution factor (D = {3, 5, 10, 20, 50, 100, ∞}. The dilution series was measured in two batches D_a = {5, 20, 100, ∞} and D_b = {3, 10, 50} in order to simulate batch effects. An inspection of figure 19 shows that our initial “guess” was wrong. In fact we can see that the cumulative intensities throughout different intensity ranges increase as a function of D until D_{100}. The MeOH blank shows lower intensities than D_{100}. Does the same apply to the peak counts per mass spectrum? In fact, figure 20 shows the exact same pattern of peak counts over D which in addition means that intensity and peak counts are in a positively correlating relationship.

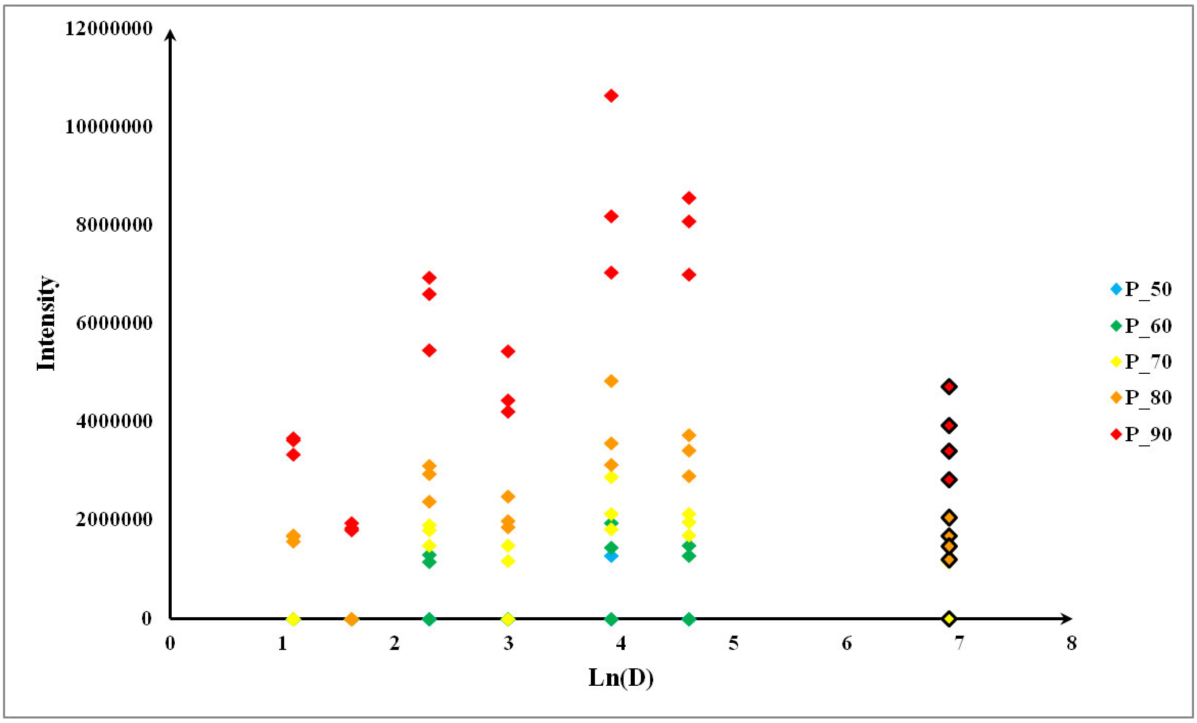


Figure 19: Intensity percentiles P = {50, 60, 70, 80, 90} over dilution factor D. The 100th percentile was excluded; the power distribution common to all mass spectral intensities made it impossible to plot the intensity profiles together.

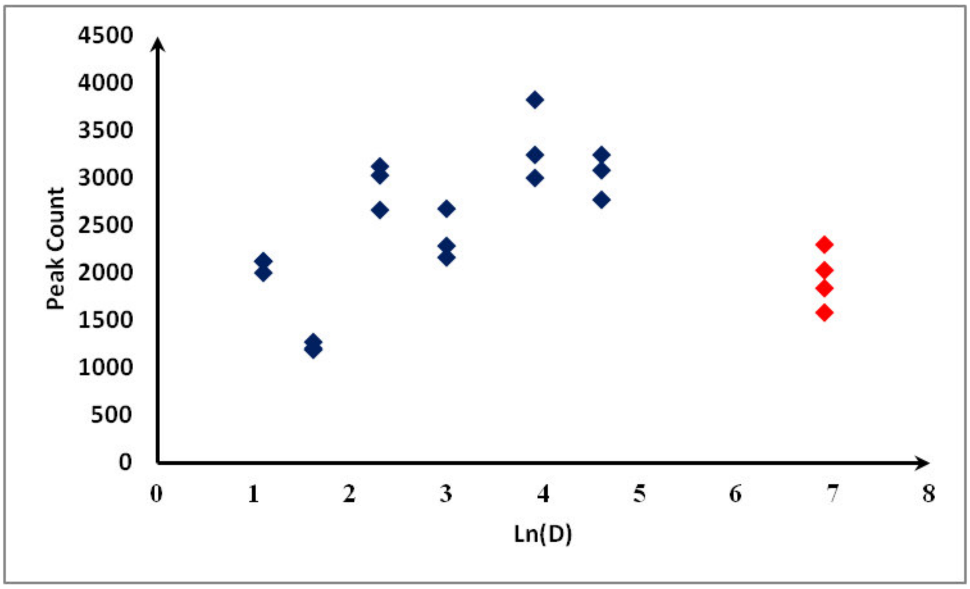


Figure 20: Peak Counts over dilution factor D

This non-intuitive finding raises the question, as to whether a higher peak magnitude in direct infusion MS indicates a lower concentration of analytes. It is necessary to test, whether this relationship applies to all masses or whether the cumulative values used for its visualization mask the true behavior of the data.

Finding the answer to the above stated questions necessitates the use of multivariate clustering methods because a) this is the exact task of multivariate methods and b) the one-by-one inspection of every mass line would prove to be tedious. In the present manuscript we predominantly use networks for analysis, because their analysis provides direct visualization of clusters and direct access towards important masses. We recapitulate; there are at least two different types of networks that can be created upon mass spectrometry data: Mass Difference Networks (MDNs) and Co-Intensity Networks (CINs). The term "Co-Intensity Network" was first coined in the doctoral thesis of Dimitrios Tziotis and is used to address any network pertaining to the quantitative information in the data as opposed to the mass difference information. In consequence, CINs include similarity networks and distance networks. In order to create CINs, the relationship between any pair of variables has to be listed in a co-intensity matrix, which is either a similarity matrix, or a distance matrix. Then it is necessary to generate a margin where CM entries are either listed as 1 if the relationship of two variables is above a threshold or listed as 0 if the relationship between the two variables is beneath this threshold. The resulting binary CM is again called adjacency matrix (AM) and it is the blueprint of the network.

The AM for the dilution series was created in the following way:

- Omit all variables which have less than 4 non-zero elements throughout the dilution row
- Center non-zero entries around their mean
- Normalize each variable on the p2-norm (Euclidean norm) to gain the normalized Matrix N
- Multiplication of N with itself according to $N*N^T$ gives CM
- We create AM by replacing the non-diagonal entries with $\cos\phi >= 0.9$ by 1 and replacing all other entries by 0.

We then visualize the Network in Gephi (Fig. 21).

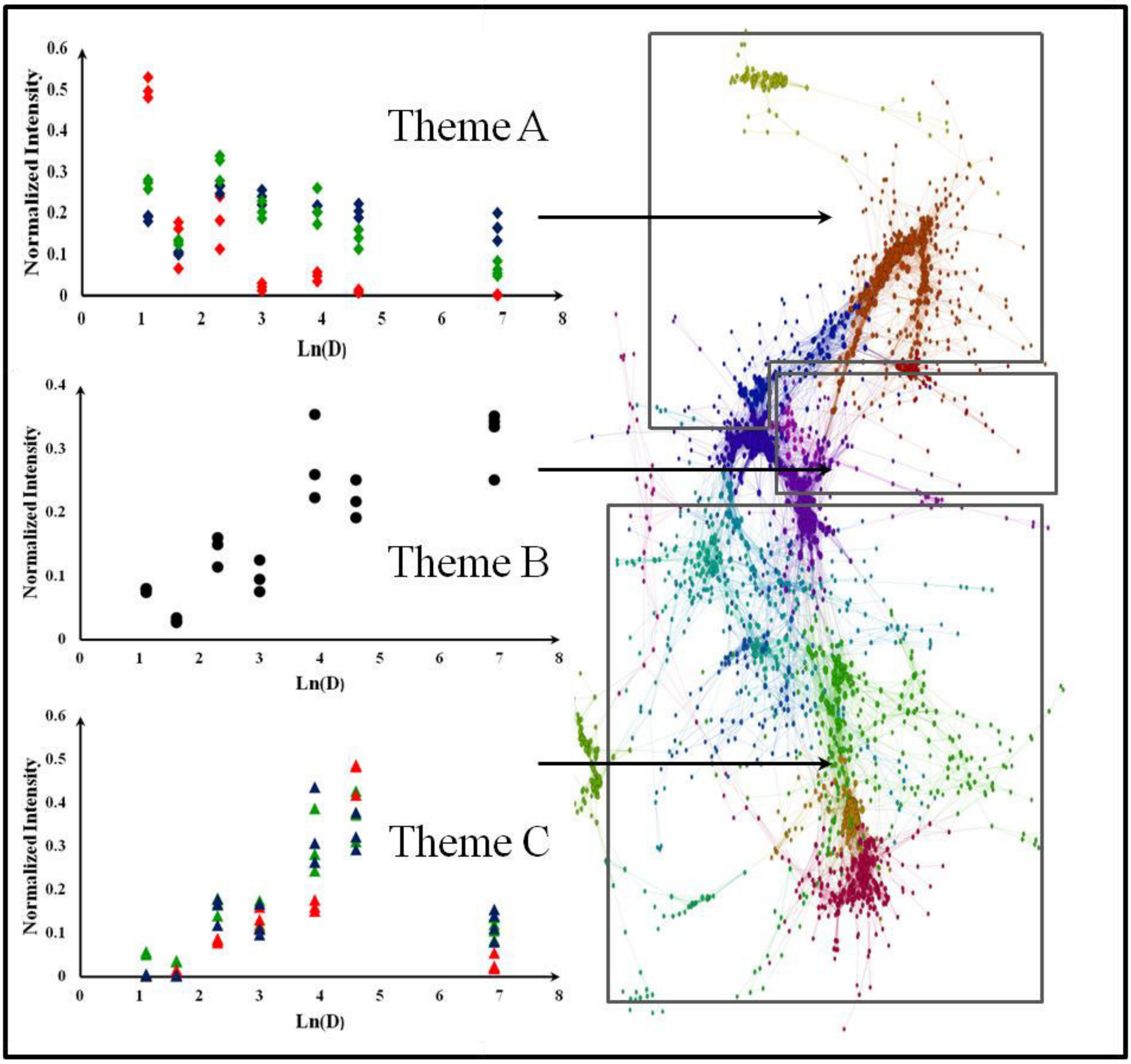


Figure 21: Co-Intensity network of dilution set. Nodes are the m/z features and edges represent relations that satisfy $\cos\varphi >= 0.9$. Nodes are colored according to modules (network sections which share more similarity among themselves than with the rest of the network). The network regions are characterized by three different themes: Theme A refers to 468 of 2178 (21.5%) intensity profiles that decrease with increasing dilution. Theme B refers to 53 of 2178 (2.4%) intensity profiles that strictly increase with increasing dilution. Theme C) refers to 1657 of 2178 (76.1%) intensity profiles that increase with increasing dilution but are barely present in 100% MeOH.

The CIN in figure 21 shows very good modularity. The intensity profiles represented by each node can coarsely be grouped into three themes. Theme A refers to profiles, which are either strictly decreasing as a function of dilution or to profiles which show a decrease from D_3 to D_5, an increase from D_5 to D_{10} and a strict decrease with dilution afterwards. Theme B refers to a small group of profiles which strictly increases with increasing dilution. Theme C refers to the majority of profiles, which increase with increasing dilution but are less abundant or barely detectable at infinite dilution (100% MeOH).

Theme A is coherent with our initial “naïve” assumption and can be taken as representatives of the EBC metabolome. In most cases signal intensities decrease at first, but increase at D_{10}. We hypothesize, that the proportion of H_2O, stemming from EBC at D_3 and D_5 destabilizes the electrospray, which indicates, that 90% MeOH (D_{10}) gives the optimal compromise between dilution and sample composition. Theme B profiles must clearly relate to the solvent used for dilution (MeOH). Theme C is the exact same pattern as shown in figure 19 and figure 20. Its dominance (76.1% of all nodes) explains why the sum of all intensities as well as each percentile profile throughout the data reflects theme C. But how can this result be interpreted in terms of dilution? The MeOH blanks were measured before, in between and after the acquisition of the dilution series and they show tremendously lower or no ion abundances for most of the intensity profiles of theme C. In-house we explain this phenomenon by the decreasing influences of theme A, i.e. by decreasing suppression due to dilution of theme A compounds. We see it as imperative to investigate the optimal sample dilution prior to any direct infusion MS experiment in order to maximize the coverage of matrix constituents.

The bi-directional nature (the mix of theme A and theme C) poses problems for data interpretation, as we cannot confidentially say, which of both themes applies to a peak in a given analytical scenario.

Let us recall, that the aim of our investigation was to come up with a strategy of data normalization, so that the EBC constituents would reflect similar ion abundances of the technical replicates, which are confounded by dilution. At this point we have to conclude, that it is not possible to normalize the data in a way, that both, theme A and theme C are corrected, because if we correct theme A by normalizing on theme A patterns would confound theme C and vice versa. We can further conclude that the use of one single dilution marker is not sufficient and that normalization techniques, which are based on representative values of data variation cannot be capable of this feat.

The correct procedure for normalization must therefore be to cluster the data and to then normalize the constituents of theme A by their representative value and to do the same for theme C constituents.

A sample set of a metabolomics study, however, is not as ideal as a dilution series. Such a set is typically composed of at least two different phenotypic groups, which may alter the proportions of theme A constituents and theme B constituents. In addition, a co-intensity network may be “confounded” by the experimental emphasis to a degree, that theme A and theme B are not immediately clusterable.

Co-intensity structure of the dilution series confounded by different pretreatment of samples

In order to develop the optimal data pre-treatment strategy we have to test, to which extent it is possible to differentiate theme A and theme B under confoundedness. We mix the dilution series with mass spectra of different SPE sample preparations of EBC. As in the previous subsection, we investigate the mass spectrometric quality of the unified dataset and remove mass traces which may invoke problems.

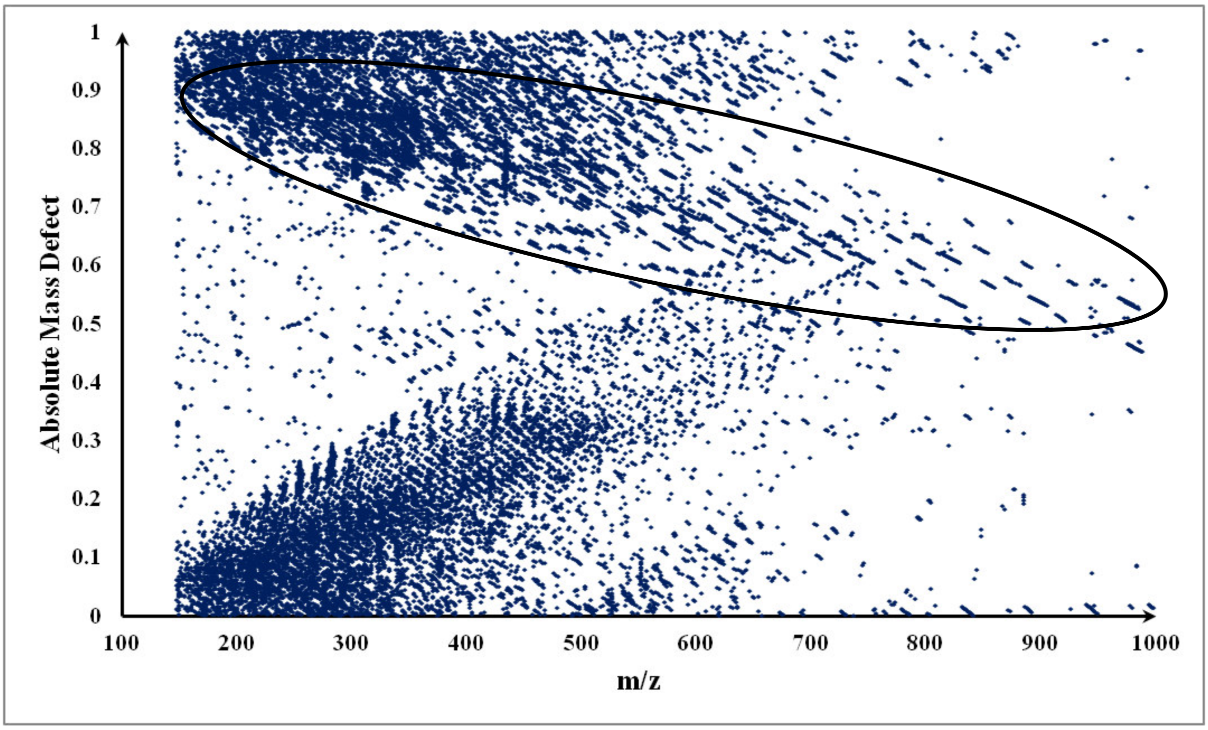


Figure 22: Absolute Mass Defect over m/z plot. The ellipsoid indicates the presence of salts.

Figure 22 shows that the addition of SPE pre-treated samples introduced a major proportion of contaminations, which are constituted of salts. The presence of salts indicates that the washing steps in the SPE preparations were not sufficient even though they were performed following the vendors protocols.

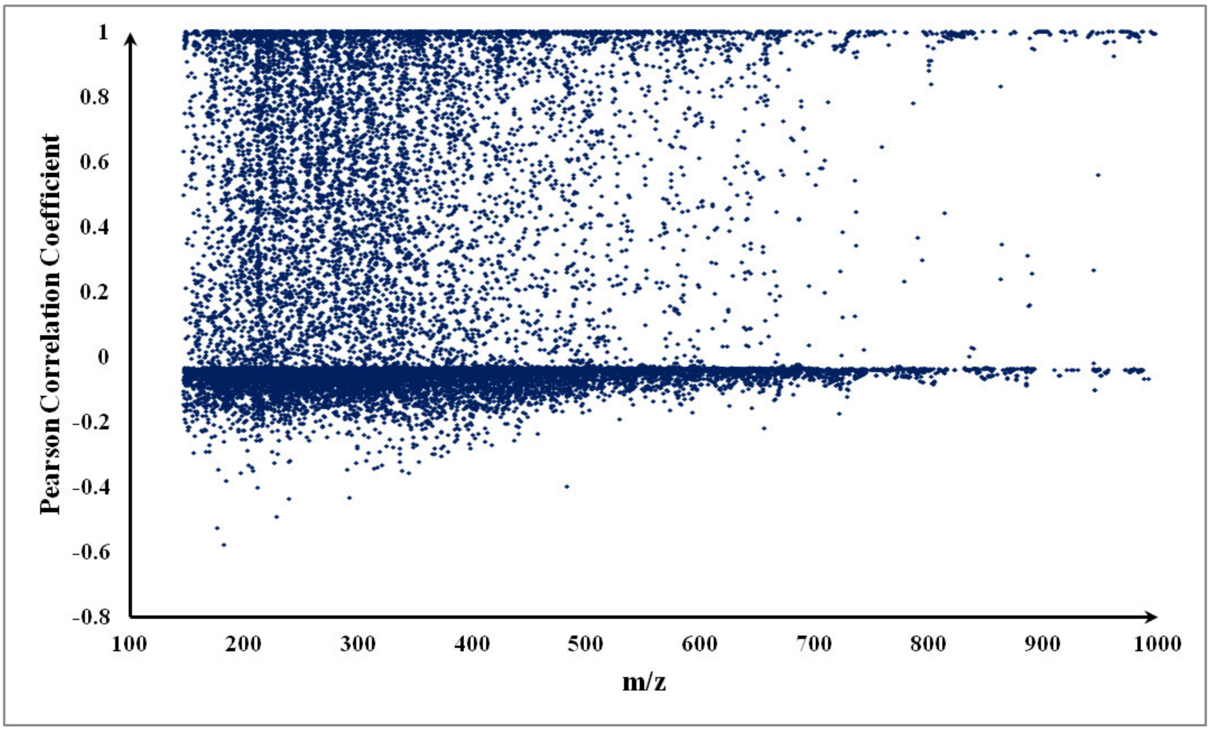


Figure 23: Running Pearson Correlation Coefficient over m/z plot.

Figure 23 indicates an increased proportion of m/z values which are dependent on their neighbors. In addition we can see a very high and non-gaussian densitiy of m/z values which are zero or smaller than zero. Figure 24 shows, that these values are associated with a low proportion of non-zero entries in the mass-sample intensity matrix.

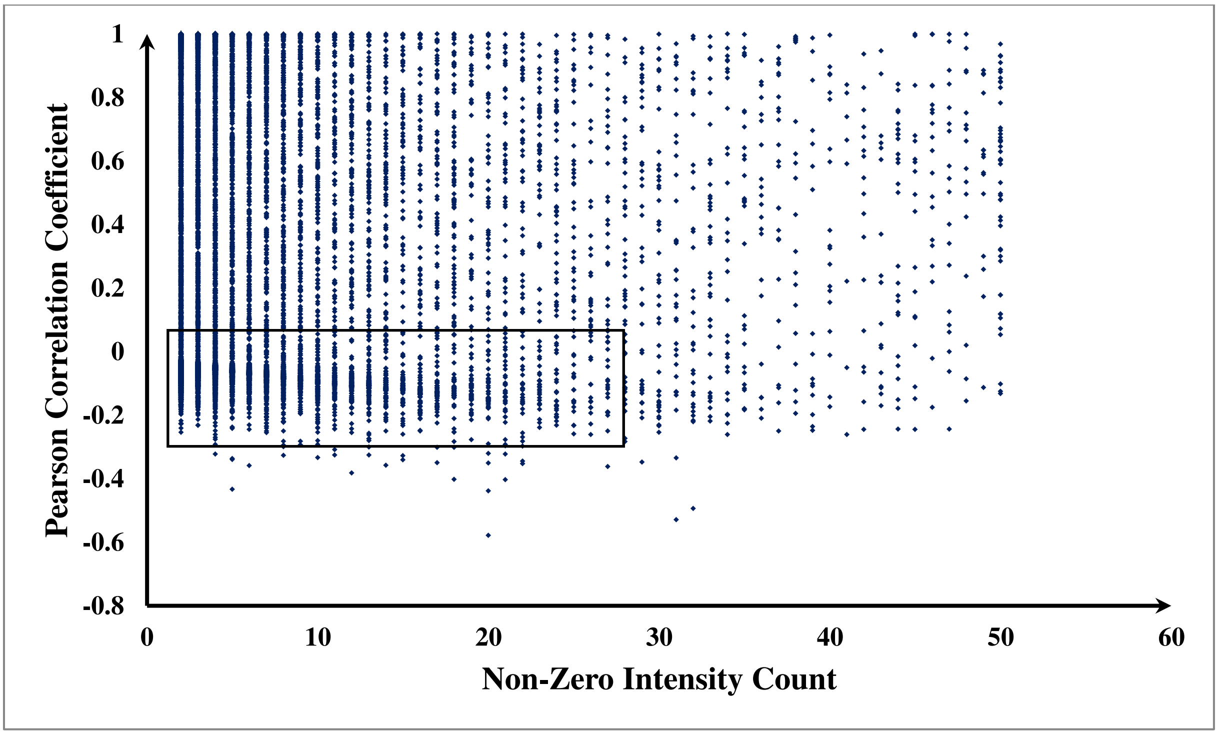


Figure 24: Running Pearson Correlation Coefficient over frequency of non-zero entries.

On the other hand side, figure 24 shows a low proportion of negative correlations which indicates acceptable calibration and alignment.

The manual inspection and filtration of the “faulty” m/z features is tedious and therefore we use Netcalc annotation in order to filter the masses. In theory – and also multiply verified in house – Netcalc excludes salts and multiple charges from the dataset.

For Netcalc annotation we use the metabolomics mass difference list, an edge formation error of 0.1 ppm and we start annotation at palmitic acid ($[C_{16}H_{32}O_2\text{-}H]^-$ ≈ 255.232954 m/z) and glucose ($[C_6H_{12}O_6\text{-}H]^-$ ≈ 179.056112 m/z). Figure 25 shows the mass spectral evaluation of the 3558 Netcalc annotated m/z values. This amount equals a proportion of annotated masses of 18% given the entire dataset. Excluding forbidden-zone compounds, the proportion of annotated masses is 30.7 %. The non-annotated remainder commonly pertains to non-CHONSP compounds and isotopologues.

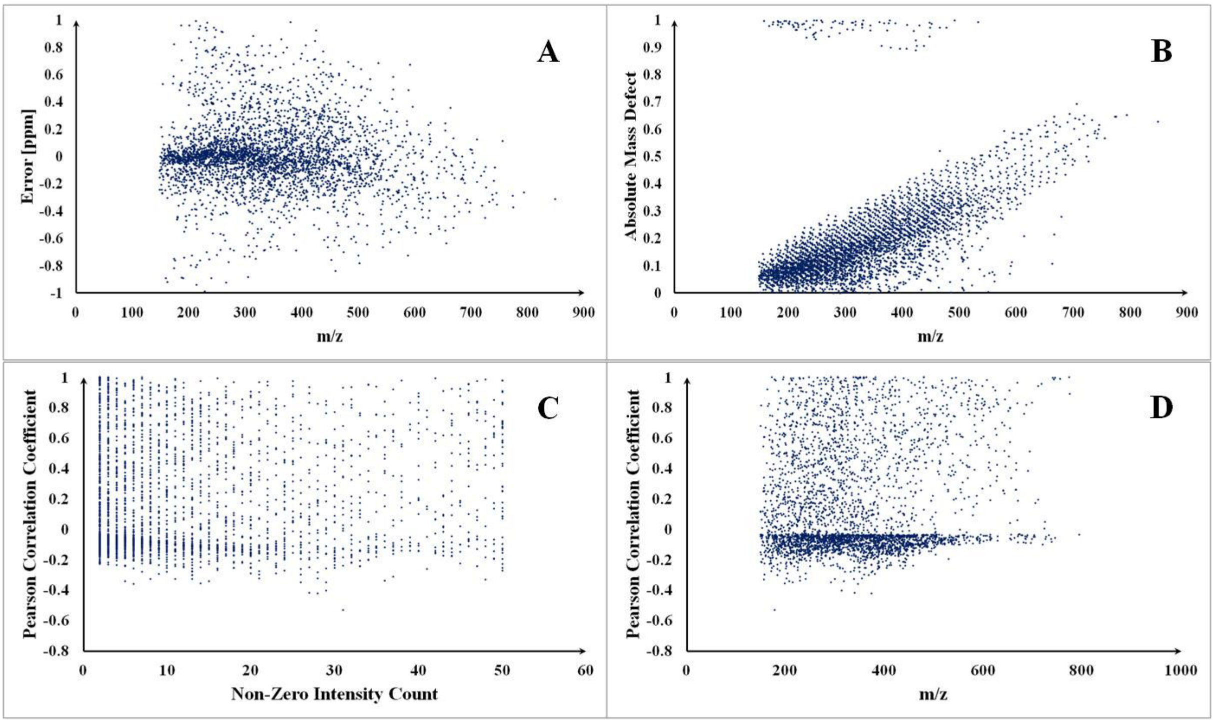


Figure 25: Characteristics of the Netcalc filtered dataset. A: The Error distribution of Netcalc annotations is well centered on zero ppm, which indicates precise and identical calibration throughout all 50 mass spectra in the set. B: In contrast to figure 22 the AMD over m/z plot is perfectly clean, which underlines the performance of Netcalc. C and D: The running Pearson Correlation Coefficient plots show, that Netcalc excluded co-linearity.

Further investigations of intensity profiles are carried out on the basis of the Netcalc annotated dataset. As elaborated above, we will use a CIN – prepared in exactly the same manner – in order to analyze the correlation structure of the data.

Other than the network in figure 21, which was one connected component covering 54% of all nodes, the current CIN shows a set of seven disconnected graph components, which altogether amount for 59.3% of all connected nodes. Each disconnected graph component addresses a different theme or pattern which exists throughout the data. All connected nodes covered 63.4% of the entire dataset.

Where the original dilution set was composed of the themes A, B and C, exclusively, the present network contained only 103 of the original 2178 features that related to the dilution series. In consequence, the entire co-intensity structure of the mixed dataset is dominated by the strong differential efficacy and specificity of the used SPE data. The rest of all nodes exclusively referred to specificity patterns of sample preparation.

Two conclusions can be drawn from these results: SPE sample preparation strongly distorts the EBC metabolome and strong variations in sample pre-treatment – e.g. unclean laboratory praxis (here simulated) – impair data quality to such an extent, that a matrix effect which is based on differential dilution cannot be corrected.

Co-intensity structure of the dilution series confounded by non-SPE samples

In order to test, whether matrix dilution is the major variable that dominates the co-intensity structure of data sets generated based on the same sample pre-treatment, we exchanged the SPE samples by samples derived from specimen number five of the HuMet study, which will be treated in the HuMet chapter. These samples were pre-treated by dilution in methanol only. The dataset passed all previously introduced tests and both partitions, the humet partition and the dilution partition, were scaled (not centered) separately in order to neutralize differences in intensity levels. In fact, even though both sets were measured within one month under the exact same experimental settings, the HuMet data exhibited much lower intensities and the median peak counts amounted to only 20% of the in house sampled EBCs. There may hence be a systematic depletion in metabolites throughout the HuMet set which is treated in the HuMet chapter. The 3183 features of the HuMet-Dilution dataset were again scaled on the Euclidean norm and non-zero intensities were centered about the mean of the non-zero intensities. It is important to center the non-zero intensities only, because otherwise weight is transferred to zero values. The resulting co-intensity network contained a connected component which encompassed 83.9% of the data, and which could be clustered into seven modularity classes. The graph component consisted of 26 modules, five of which (16.7%) were fully related to the HuMet set and the dilution set.

The module-wise sums of the scaled data over the samples revealed distinct patterns, which exclusively represented different characteristics of theme C. Themes A and theme B (the

MeOH matrix) did not associate with the co-intensity structure of the HuMet set.

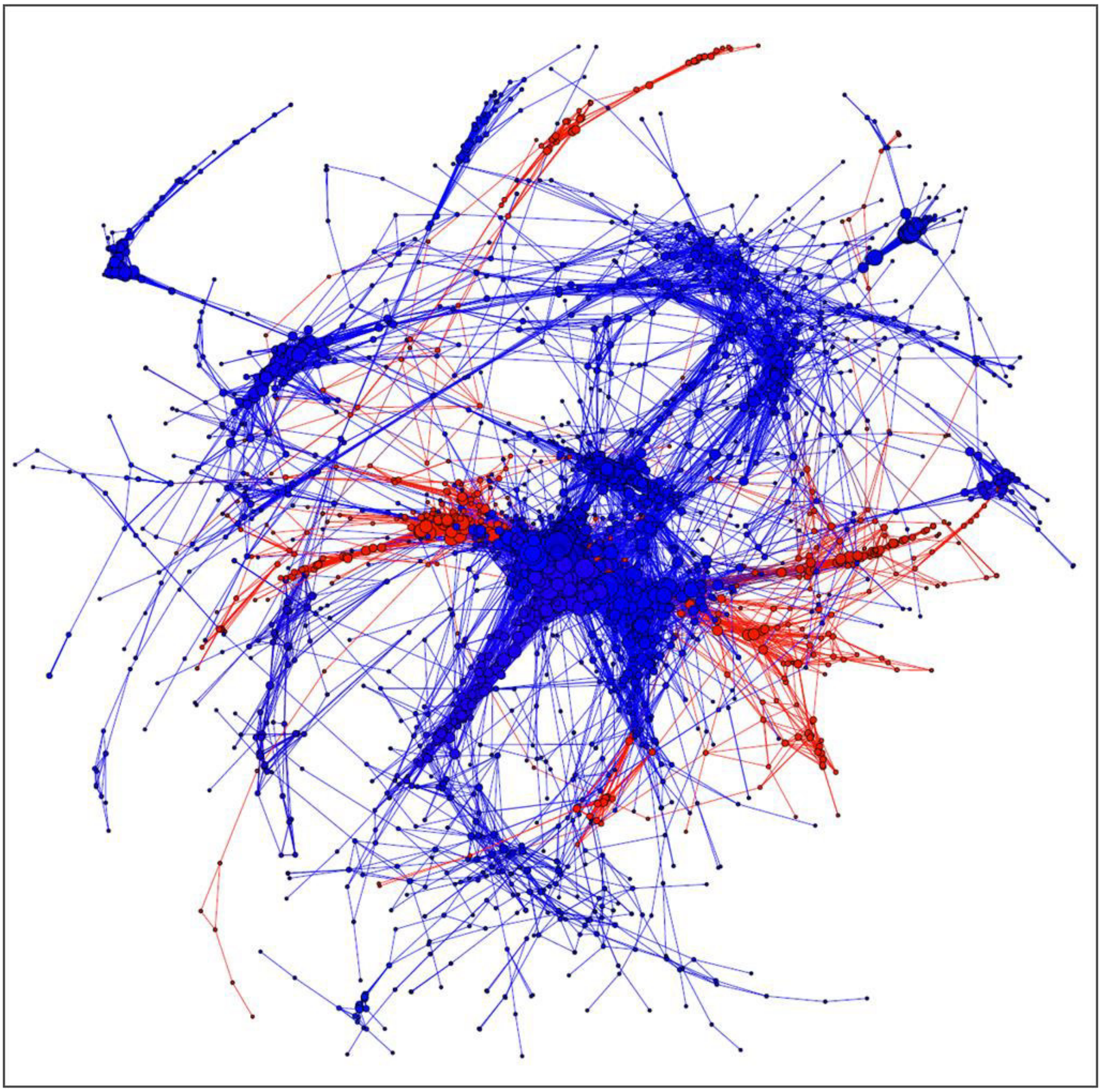


Figure 26: CIN of the HuMet-Dilution set. Nodes of the HuMet partition are red. The HuMet modules are well associated to the main dilution topic, which is theme C.

The dilution themes of the HuMet-dilution modules are associated with theme C; however, different sub-themes are addressed.

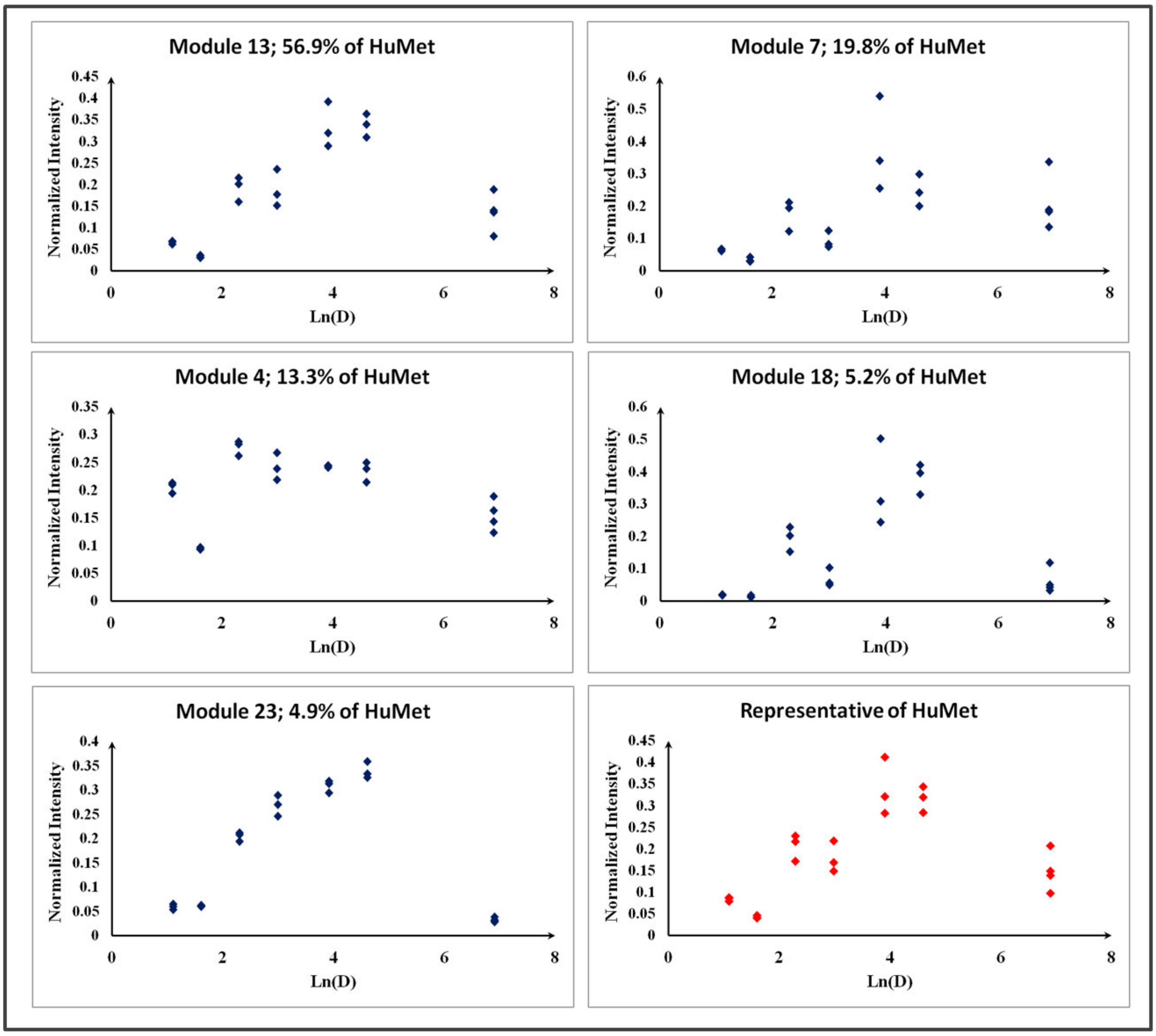


Figure 27: Different Modules of theme C that clustered in the co-intensity network and representative sum of scaled HuMet data.

What differentiates the sub-patterns of theme C is the magnitude by which the respective features reacted to dilution. The fine structure of theme C was supported because the centering applied prior to network creation was based on the non-zero entries only. Including zero entries would have dulled the precision of the picture.

Even though the patterns appear to be very similar, they express differences. The visual similarity of the patterns may imply that normalization on the representative (which is the Manhattan norm, Taxi Cab norm or p1-norm) neutralizes the dilution behavior. Principally all samples of the dilution series are technical replicates and normalization is finally intended to assign the same intensity to all dilution stages (based on a norm which is a representative of intrinsic dilution behavior). At this point we have to envision, that theme C implies that higher intensity equals lower concentration! In consequence, non-normalized or incorrectly normalized DI-ICR-FT-MS data, which implies statistically significant up-regulation of a marker candidate in a phenotype of interest may in reality be a marker of significant down-regulation. The aim of normalization must therefore either be to neutralize the dilution effect (matrix effect) or at least to make sure that a lower intensity really correlates to lower concentration.

What happens, if we normalize the HuMet modules on the representative?

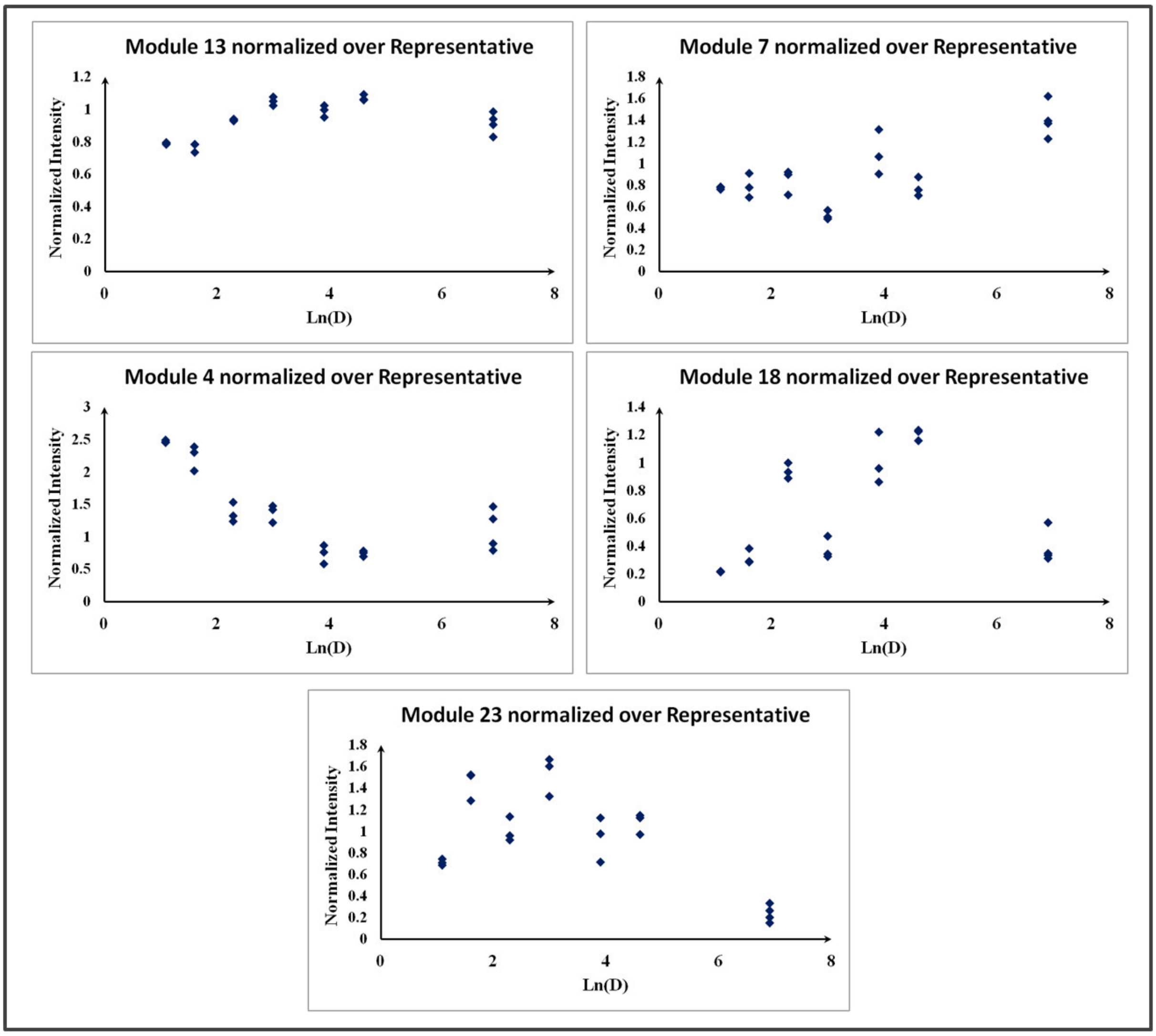


Figure 28: HuMet modules normalized on representative data profile.

We can see that the normalization onto the representative data profile works well for modules 13 and 7. The other modules are distorted; module 4 shows a trend which is conform to sample dilution, but the last quadruple (infinite dilution), is over-estimated. We remember that features, which appear at this dilution stage mostly come from carry over, which is why the normalization result for module 13 is acceptable. In the rest of the data the dilution response is either distorted (not neutralized) or the infinite dilution/carry over section is over-represented.

In conclusion we can state, that it is possible to define modules of common matrix effect and that it is appropriate to normalize the respective data onto ‘personalized’ matrix patterns.
The definition of modules is mathematical and rather strict. It might be – and the more in networks of large degree like the present co-intensity networks – that a specific feature has 49 connections to theme X and 50 connections to the theme Y and that its normalization is then based on theme X exclusively. As we now assume, that the primary impact onto co-intensity structure comes from dilution, we can correct the features on their neighbors instead of correcting them for their modules. Of course, we have to bear in mind that both, module sizes and neighborhood sizes may vary, which gives a different scaling after normalization; so we need to rescale the normalized data.

3.4.2 Test of normalization strategy

Based on the previous sections, we developed the following normalization strategy:

1) Creation of a co-intensity network
2) Calculation of a ‘taylor made’ norm for a given feature, based on the features which are adjacent in the co-intensity network

We performed this strategy on the dilution set and compared it to the application of the Manhattan norm. It is not important, which norm is used in this case, because the data was scaled before. Scaling cancels out the power relation of intensities over the mass spectrum. Most normalization techniques produce similar results in that one single measure of magnitude is used as a norm. Which technique should be used, is largely a matter of preference and a matter of centering or not centering prior to normalization. However, it should always be born in mind that metabolomics data is often compromised by missingness, which makes p-norms the more appropriate measure.

Remark: Data mining tools commonly offer scaling and/or normalization procedures for data pre-treatment. However, the reason why these techniques should be applied, which effect they have and in which order they should be applied is often left unclear. In principal, data first has to be scaled in order to neutralize power structure. Normalization would be confounded if the power structure would not be dealt with before. Normalization changes the relative intensities of features throughout the samples. Common normalizations do not require a re-scaling. Our CIN based normalization, however, requires re-scaling after normalization, since the magnitude of the norm depends on the connectivity of a feature in the network.

In general, data should first be scaled (not centered) prior to normalization. *End of Remark.*

Assuming, that normalization diminishes the spread of the data we used the adjusted inter quartile ranges (IQRs) in order to compare both normalization strategies. What we call the adjusted IQR is the IQR over the non-zero intensities. In order to compare, whether the IQR for the CIN normalization is generally lower than the IQR of Manhattan normalization, we have to make a plot of the ratio $IQR_{CIN}/IQR_{Manhattan}$.

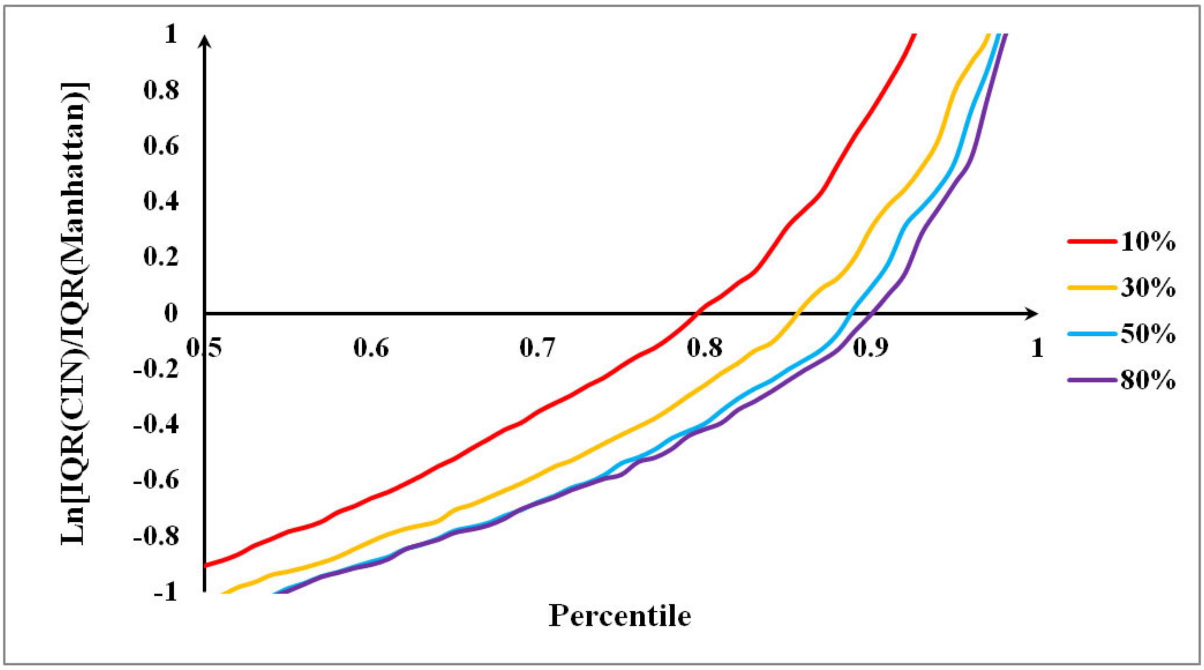


Figure 29: Zoom into the $\ln(IQR_{CIN}/IQR_{Manhattan})$ percentile plot of the dilution dataset. Different cutoffs at 10%, 30%, 50% and 80% non-zero intensity frequency. Y-axis values < 0 indicate a smaller IQR_{CIN} than $IQR_{Manhattan}$ and the point of interception of the X-axis indicates the percentage of features for which $Y \leq 0$ is true.

The CIN normalization approach is superior to the Manhattan norm throughout a wide range of non-intensity frequency cutoffs. The higher the frequency cutoff is, the better is the normalization. In consequence, there is a relation between missingness and normalization result. This outcome is to be expected, since the missingness structure of a dilution dataset is given to be systematic. The direct consequence is that sums over samples with high missingness, e.g. infinite dilution, D_3 or D_5 are smaller, which leads to a smaller norm and finally to an overestimation of present non-zero entries.

For this reason it is important to secure ‘missingness at random’ in a given dataset.

3.5 Conclusion

We have demonstrated, that features react in different ways upon matrix effects and we have drawn the conclusion that normalization on a single norm must fail to normalize every feature justly. We have shown that SPE pre-treated EBCs contain a large amount of contamination and that SPE patterns are fundamentally different from dilution patterns. This indicates that minimal sample pre-treatment is the better strategy for EBC, which has low amounts of analytes to begin with.

We have indicated that the major part of correlation in the data comes from bias. This is not necessarily new, since normalization and scaling would be unnecessary if this was not the case. In consequence, we could develop a co-intensity network based normalization approach, which is more capable to neutralize bias than normalization approaches, which are based on a single norm. In addition we found out, that – apart from data cleaning – it is necessary to first make sure that the missingness structure in the data is at random, to then scale the data (without centering), to then normalize the data and, if necessary, to perform a final scaling (with or without centering).

4. Data Analysis 2: The Gauting Study

The Gauting Study [Möller, W., et al., 2009] was intended to show that EBC is a proper analytical matrix for the diagnosis of chronic obstructive pulmonary disease. EBC was sampled from smokers, non-smokers and ex-smokers, which had COPD. The samples were collected using the EcoScreen 2 sampler, which separates alveolar and bronchial breath streams. Samples were measured using the NanoMate roboter.

In the present treatment of the Gauting study, COPD samples had to be omitted because they were too few (N = 5) and they were cortisol treated, which enforced an emphasis on sterol metabolism in early analyses.

A direct comparison of alveolar and bronchial EBC was intended originally. However, the implementation of a 'pre-annotated' reference sample for internal calibration has revealed, that the error distribution of the bronchial samples were strongly non-linear. Specifically, an intense m/z peak cluster at m/z ≈ 400. Marshall et al. demonstrated that the linearity of error distributions in ICR-FT-MS depends on peak intensities. Previous analyses of the Gauting set (calibrated on solvent impurities) revealed that a main proportion of variables that separated the alveolar samples from the bronchial samples were characteristically found at m/z ≈ 400. The use of a large calibration list, which was able to properly resolve the error distribution, revealed that the largest difference between alveolar error distributions and bronchial error distributions was at m/z ≈ 400, which resulted in misalignment of the spectra. In future approaches more adaptive regression models like LOWESS [Cleveland, W.S., 1979] in conjunction to mass difference networking will be applied outside of Bruker Data Analysis in order to overcome such problems. In consequence, no direct comparison of alveolar and bronchial samples is performed here.

After the unification of the alveolar smoker samples (N=12) and non-smoker samples (N=13), consequent elimination of m/z values with a frequency < 3 and Netcalc annotation we gained a dataset of 3711 m/z values over 25 samples.

4.1 Analysis of data structure

We perform a first analysis of the data structure by using (multivariate) linear algebra. Similar to PCA, we want to use eigenvectors as representatives of data structure. The following steps need to be performed:

1) Create dataset D
    a. Raw dataset
    b. Normalize all rows on their Euclidean norm
    c. Normalize all columns on their Euclidean norm
2) Create the coincidence matrices of D by calculating **CM** = $\mathbf{D*D^T}$
3) Calculate the eigenvectors of **CM**
4) Select the eigenvectors with the largest eigenvalues (e.g. usually the first three eigenvalues) and multiply them with all mass spectra
5) Plot the Eigenvectors against the frequency
6) Plot the resulting loadings over the samples

Plotting the first eigenvectors of the Gauting set versus the m/z signal frequency does not allow for visual interpretation, since the eigenvectors reflect the power structure of the data.

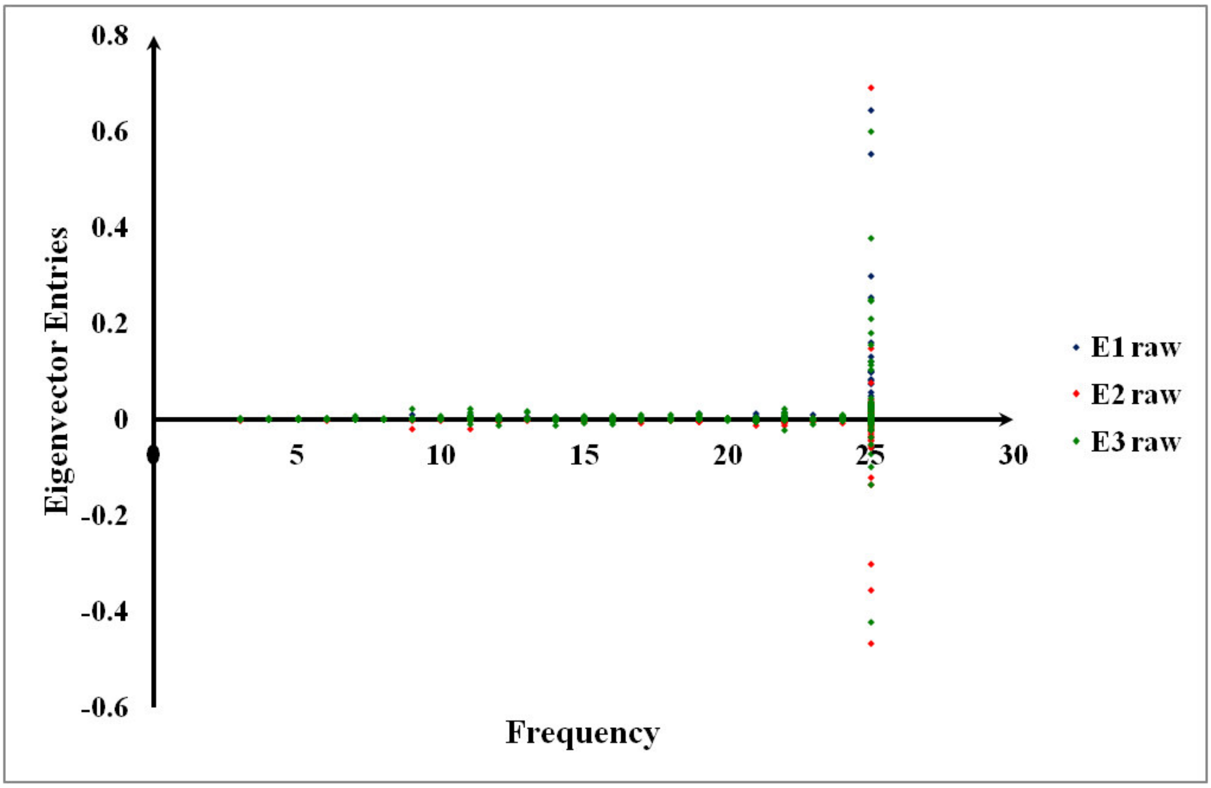


Figure 30: Eigenvector entries generated on the raw Gauting dataset plotted over frequency of m/z signals. The frequency of the largest positive and negative eigenvector entries is the maximum frequency, which indicates large intensities at large frequencies and the presence of different patterns in the data.

Normalization of the rows of the dataset onto their Euclidean results in more readable plots because this process cancels out the power distribution over the mass spectra.

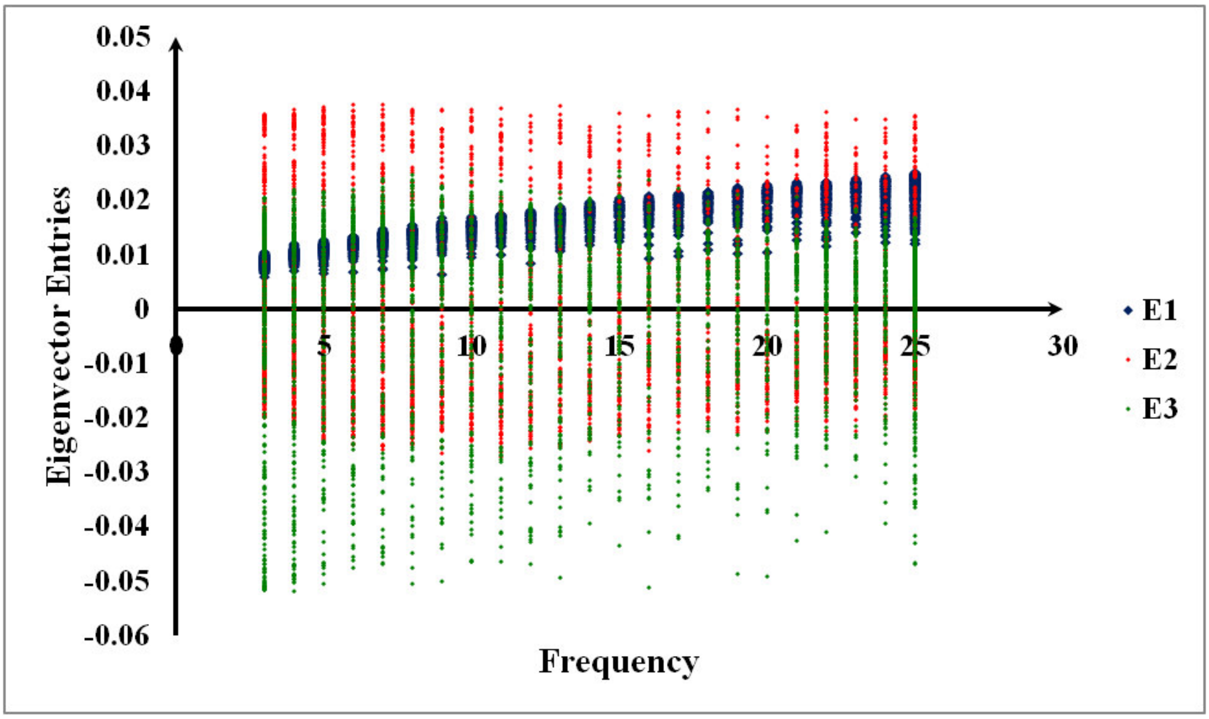


Figure 31: Eigenvector entries generated after normalization of rows, plotted over frequencies of m/z signals. The power distribution is cancelled out. The first eigenvector (E1), covering 43% of all eigenvalues, increases with increasing frequency. The other eigenvectors are independent of the frequency.

The first eigenvector shows a behavior that is linearly dependent on frequency. This behavior is to be expected, since an m/z value with many entries has a higher chance to coincide with other m/z values over the samples than features with fewer entries. If higher eigenvector entries were to cause large eigenvector entries, we would expect a strong bias. The other Eigenvectors behave entirely independent from m/z frequency. We can assume that the data structure does not contain strong binary biases.

Performing the same sequence onto the column-normalized dataset, results in the same plot as performing it on the raw dataset.

Plotting the sum of sample intensities over the peak count per sample indicates good data quality because the peak sum is independent of the peak count. If the sum of intensities would correlate with the peak count, a matrix effect would be indicated. In fact, the present scenario suggests, that ionization using the Nanomate robot was close to complete.

The analysis of the data structure indicates an optimal situation for the normalization of the data. Since the peak count is very constant, it is possible to perform robust normalizations like normalization of the inter-quartile range (IQR) or transforming each sample to equal median.

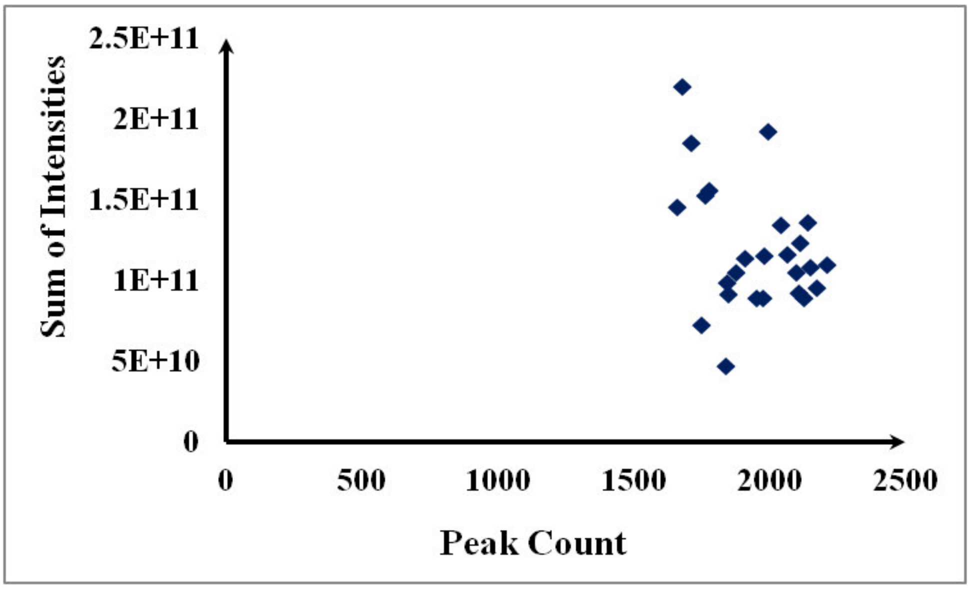


Figure 32: Sum of intensities over peak count.

The advantage of robust normalizations is that they are not leveraged by outliers. If the previous analyses had been dependent on the missingness structure and peak counts would have varied strongly, robust measures like the median or the inter quartile range would have been biased themselves (this goes for any other normalization technique as well).

In consequence, we first normalized the mass spectra on their IQR and then normalized the m/z values on their IQR. However, because of differential missingness we used a modified IQR, which is only performed on the non-zero entries.

4.2 Extraction of potential Surrogate Markers

4.2.1 Univariate analysis

Prior to multivariate data analysis we performed a simple two-sided T-test with the null hypothesis, that the intensities in non-smokers and in smokers have equal mean given different variance.

The test revealed 126 m/z values with $p < 0.05$ which had higher intensities in non-smokers and 43 m/z values with $p < 0.05$ which had higher intensities in non-smokers. Table 2 summarizes these data.

Univariate discriminative features pertain to the differential abundances of analytical signals over phenotypic groups of interest. They are in the focus of targeted analysis and the development of quantitative detection techniques allows for their direct evaluation in raw spectra.

Univariate discriminative features as they are produced here cannot be guaranteed to be found in raw spectra. We recall that the sum of intensities varied independently from the peak count and that normalization was necessary. This circumstance highlights the necessity to separate targeted and non-targeted metabolomics, because the latter, may find surrogate marker candidates and the prior may improve their quantitative detection and identification.

4.2.2 Multivariate Markers

For the extraction of multivariate markers we performed a simple PCA based on the Nth dimension (the feature dimension) of the dataset. The resulting principal components (3711 eigenvectors) were then multiplied with each sample in order to gain their loadings into the samples.

Then we performed a series of t-tests on the eigenvectors in order to extract the eigenvectors, which generated significantly differentiating loadings into the smoker and non-smoker samples.

We then extracted the three most differentiating eigenvectors, which caused loadings with the following p-values:

Table 2: Eigenvectors with respective p-values

Eigenvector number	p-value	Group	Number of associated features	Resulting p-value
205	0.0002	S	82	0.0009
608	0.0003	NS	46	0.0001
625	0.0005	S	20	0.12

The number of the eigenvector indicates its magnitude. The first eigenvector in the table is the 205th largest eigenvector of 3711 eigenvectors. We see that it is not always the first three components that contain the desired differentiation of the data. Afterall, PCA is merely optimized to reconstruct the dataset, not to find significant differentiations.

In order to find out, which masses were responsible for the direction of the eigenvectors, we wrote a short algorithm that tested whether the deletion of a variable together with its respective eigenvector entry would impair the shape of the eigenvector loadings.

In results, we gained a set of 82 features that associated strongly with eigenvector 205, 46 features for eigenvector 608 and 20 features for eigenvector 625. While the selected masses improved the p-value of eigenvector 608, they slightly impaired the p-value of eigenvector 205 and strongly impaired the p-values for eigenvector 625.

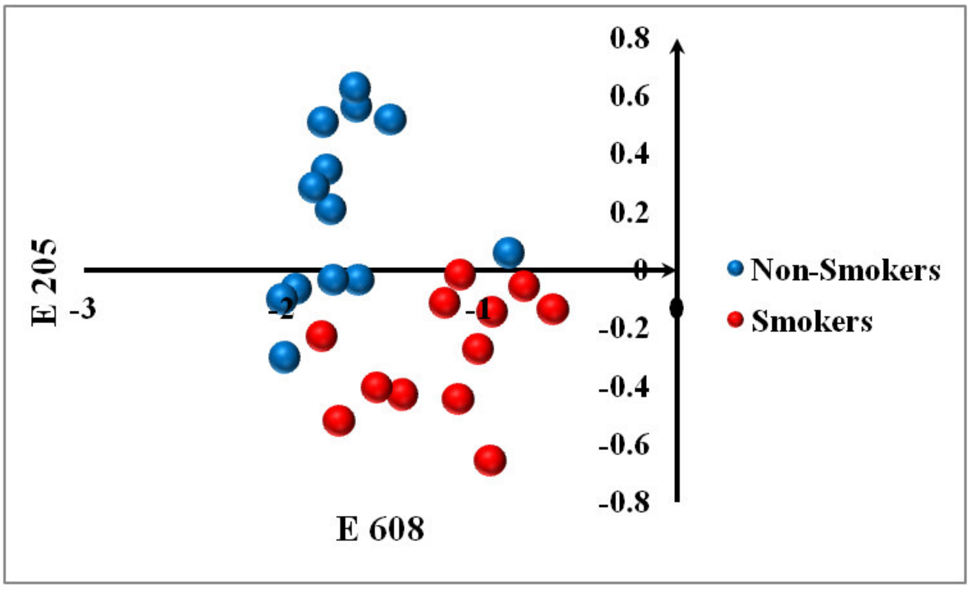


Figure 33: PCA of the Gauting dataset differentiating smokers versus non-smokers over the eigenvector E205 and E608.

An interesting finding is that the single masses, which lead to the given separation, were mostly not addressed in the univariate datamining, i.e. they do not differentiate the groups as a single feature, but they do as a group. Most of the found markers had no significant t-test result.

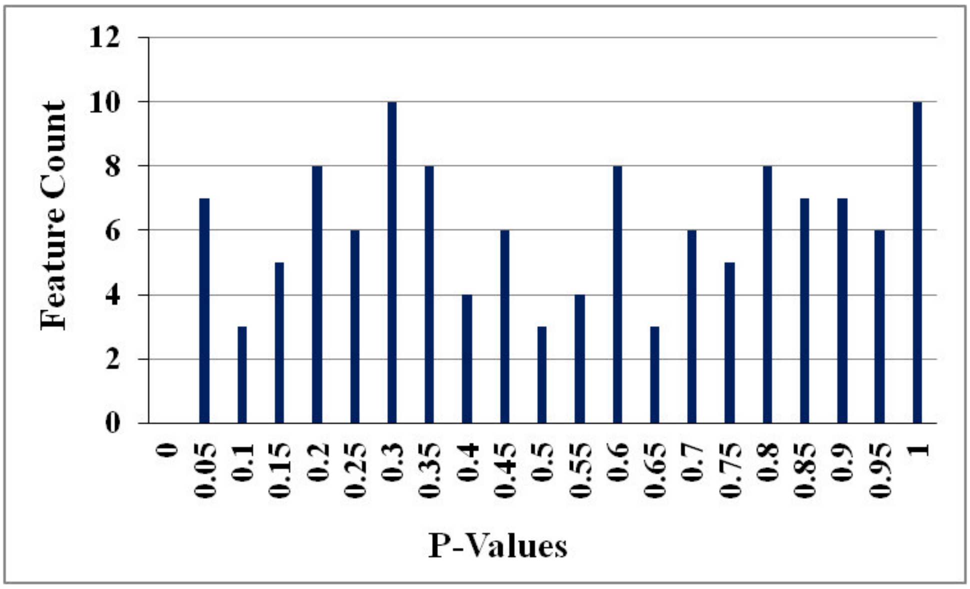


Figure 34: P-value distribution over the multivariately discriminating features

It is not the significance of a single feature that makes the multivariate marker; it is their mode of co-occurrence. The eigenvectors show invariable directions in the data and can be used as a model vector. Multiplication of the respective data with the model vector, shows how much aligned the differentiating features are with its direction. By extracting the minimum amount of features that is necessary for a good alignment with this direction gives us the respective markers. Non-smoker markers co-occur most in the non-smoker samples and smoker markers co-occur most in smoker samples.

To use multivariate techniques for data analysis and yielding a good separation does per se not imply differentiating levels of single markers. This fact raises problems for the definition of a surrogate marker, since a given device would have to be able to detect all markers of interest. However, quantitative measurements in the scope of targeted metabolomics may enable the direct extraction of representative geometries among the markers which make up the desired pattern.

4.2.3 MassTrix annotation of Markers

Database matching is at some point indispensible for non-targeted metabolomics research, because hits in metabolic databases may enable biochemical interpretation of the data. Even though there is a large number of isomers for a given exact formula annotation, data that supports the existence of a given isomer in a given sample type may indicate the most probable feature identity.

As discussed above, database matching on experimental data is problematic, since there may be too many hits that are within a specified error range but that are in no relation to the error distribution of the data. For this reason we matched the Netcalc annotations using MassTRIX at 0.1 ppm. MassTRIX annotated 431 (11.6%) of the uploaded 3711 Netcalc results. The previous data mining approaches yielded 270 marker candidates of which 30 features (11.1 %) were annotated by MassTRIX.

Table 3: MassTRIX annotations of marker candidates for Markers.

Marker Count	Name
S	5-propylideneisolongifolane
S	Prostaglandin H2
S	3-Hydroxy-9-hexadecenoylcarnitine [cation] ([M+H]+)
S	Hexanoylcarnitine
S	Xanthan
S	Oblongolide ([M+H]+)
S	Hexadecenal
S	3-methyl-tetradecanedioic acid [Dicarboxylic acids [FA0117]] ([M+H]+)
S	Retinal
S	4,8 dimethylnonanoyl carnitine
NS	Isodomedin ([M+H]+)
NS	5-O-Methylembelin ([M+H]+)
NS	Nicotianamine ([M+H]+)
NS	Hydnocarpic acid ([M+H]+)
NS	4-keto pentadecanoic acid
NS	6-Oxabicyclo[3.1.0]hexane-2-undecanoic acid methyl ester ([M+H]+)
NS	O-Decanoyl-L-carnitine ([M+H]+)
NS	L-Rhamnose
NS	6-endo-Hydroxycineole ([M+H]+)
NS	Toluene-4-sulfonate
NS	Heptanoylcarnitine [cation] ([M+H]+)
NS	12-trans-Hydroxy juvenile hormone III ([M+H]+)
NS	Butoctamide hydrogen succinate
NS	(-)-Menthyl O-beta-D-glucoside ([M+H]+)
NS	Valproic acid glucuronide (see KEGG C03033)
NS	Farnesylcysteine
NS	S-Formylglutathione ([M+H]+)
NS	8-Epiiridotrial glucoside
NS	FPL64176
NS	Pseudoaconitine ([M+H]+)

The smoker marker candidates prostglandin H2 and hexadecanal are known to be produced in response to oxidative stress. 4,8-dimethylnonanoyl carnitine and 3-methyltetradecanoic acid may be a results of oxidative stress as well. Both marker groups encompass carnitines. The involvement of S-formylglutathione into non-smoker biochemistry may indicate a working antioxidative mechanism and toluene-4-sulfonate may indicate an intact detoxification mechanism. Such detoxification mechanisms are closely linked to systemic metabolism. Farnesylcysteine is a mixture of an isoprene and cysteine. Isoprenes were frequently reported to be constituents in exhaled breath but a specific role cannot be assigned at this point.

While some annotations comply with literature resources, no larger picture pertaining to the differences between non-smokers and smokers can be compiled. This scenario – the under annotation of non-targeted marker candidates – is a common problem to non-targeted metabolomics; especially if there is limited knowledge about an analytical matrix such as EBC. In consequence, it is necessary to apply different approaches for data interpretation

4.3 Development of appropriate Workflows for the extraction and Interpretation of Surrogate Markers

Up until this point, it was possible to optimize putative annotation of mass spectral signals and to extract relevant information in a rationally understandable manner. The next step in metabolome investigations is the interpretation of results.

There is a plethora of information stored in databases, which store metabolomics, proteomics and genomics information of central metabolism, metabolism of specific plant species, metabolism in cells, mouse model systems, human plasma and urine, tissue extracts and many more. Still, in most data sets only 10% of all features intersect these stored data – in some standard model systems more and in some less.

The intersection of cleaned EBC data with the data stored in the MassTRIX server is at 5% and stays below 10% when matching against HMDB. From the biochemical point of view EBC is a black box. Consequently – at a pre-targeted stage – interpretation must largely be data driven and it is necessary to develop techniques, which lead to an objective, chemical circumscription of processes underlying the data. Where missingness of information is an uncomfortable scenario for the biochemist of physiologist, it is a common scenario in the analysis of natural organic matter (NOM). Having no pathway, protein sequence or gene list at hand the NOM analyst routinely uses the following techniques in order to devise a holistic description of analytical/experimental scenarios:

- Analysis of elemental ratios (like O/C or H/C) which are often displayed in a van Krevelen plot.
- Comparison of the density of population of specific van Krevelen regions between different samples.
- Elemental ratio over m/z plots. Elemental ratio over m/z plots give insight into the continuity of chemical processes over the m/z range, and well defined structures imply the existence of homologous series.
- Kendrick Mass Defect (KMD) plots [Kendrick, E., 1963]. Here masses are scaled

according to the proportion of the nominal mass of a composition over the exact mass of a composition. The mass defects of scaled masses plotted over the scaled masses themselves reveal homologous series in the data. The continuity and length of homologous series in addition to their differences from sample to sample give insight into structured chemical processes which are characteristic for the data.

- Analysis of the aromaticity index (AI) [Koch, B.P. and Dittmar, T., 2006] and double bond equivalent (DBE) [Pellegrin, V., 1983] gives information on the redox state of a system and on the source of substances (e.g. aromatic systems like humic acids at large AI and DBE, aliphatic – mostly anthropological petrol impacts – at low AI and DBE).

Data analysis with these tools at hand enables the observation and interpretation of compositional and chemical shifts – elemental fluxes – in mostly stochastic non-steady state environments. The main focus of these techniques is on the isolation of homologous series in conjunction to differences in abundance observed throughout the data.

The use of the same tools for metabolome analysis is largely inappropriate, since the metabolome is discontinuous in terms of CH_2, CO_2, N, NH, NH_2, O, HSO_3, HPO_3 series.

Nonetheless such type of analysis is possible on metabolome data owing to Netcalc and mass difference networking in general.

The concept of Netcalc and the first Netcalc algorithm were laid out and applied for NOM analysis. Mass difference networking itself is the multi-dimensional mapping of homologous series. Graph theory provides a set of tools and concepts, which can be used for the demarcation of important network structures. Most observed and published graphs (networks) exhibit a power law structure, sometimes a scale free structure, in terms of connectivity. As introduced before, such distributions over a sorted set of events indicate a small probability for events to occur at one end of the distribution and a large probability for events to occur towards the opposite end of the distribution. In graph theory this concept translates into a low probability to find highly connected nodes and a large probability to find poorly connected nodes.

Given such a topology, nodes in a graph are postulated to adhere to different degrees of importance, to different roles. Networks of power topology often develop a so called community structure which is expressed in the formation of modules. A module is a network region that is more densely connected within its elements than to the reset of the network. Analogously, members of a module are interpreted as being more similar to one another than to the rest of the graph.

Based on connectivity and modularity, Guimerá devised a classification of roles which nodes can take on. Nodes can be classified into 'ultra-peripheral' nodes (low connectivity within the module and no connection towards another module), peripheral nodes (nodes with a majority of its connections being within their module), non-hub connector nodes (many links to other modules), non-hub kinless nodes (nodes with randomly distributed module specificity), provincial hubs (strongly connected nodes with most of their connections within their module; they are module representatives), connector hubs (largely connected nodes with many links to other modules) and kinless hubs (strongly connected nodes, with randomly distributed modules specificity) [Guimerà, R., et al. 2005].

Depending in the context of a network, these roles can be indicative for a node's importance for network structure or representativeness for their module. Other indicators of a node's importance are the clustering coefficient or betweenness centrality. Applying such measures on a mass difference network the expectation is, that they allow immediate information on metabolic pathways or chemical processes; this was hypothesized by Breitling [Breitling, R., et al. 2006]. This, however, is only true for non-random networks.

The reconstruction of a mass difference network on the theoretical exact masses of all Netcalc annotations is inherently "error-free" and therefore represents the entire stoichiometric relationships within a given dataset. Analyzing the topology of such networks reveals that they do not have the required network topology for network analysis. Instead of a power distribution, theoretical mass difference networks have a strong tendency to be random with an unusually large amount of highly connected nodes. The following question arises: "Why are almost all published networks reported to have power or scaling distributions, but theoretical mass difference networks do not?" If an answer to this question can be found, it should be possible to correct the network topology so as to give a power distribution which in consequence allows for network interpretation. An analysis of commonly applied networks versus mass difference networks reveals important differences: commonly published networks are correlation networks, interaction networks, social networks, studies on the World Wide Web and so on. Their criteria for edge formation are either binary (do genes correlate? Yes or no! Do proteins interact? Yes or No!), or they are constructed by man like the metabolic pathway as presented by Guimerá or like the World Wide Web. They all have only one criterion for edge formation and are supervised.

The creation of mass difference networks using the above developed metabolic REMD list includes 176 criteria overall.

Hypothesis: Network topology is power distributed or scale free given a small set of REMDs and gradually becomes random as the number of applied REMDs increases.
Figure 35 shows this relationship based on the smokers dataset.

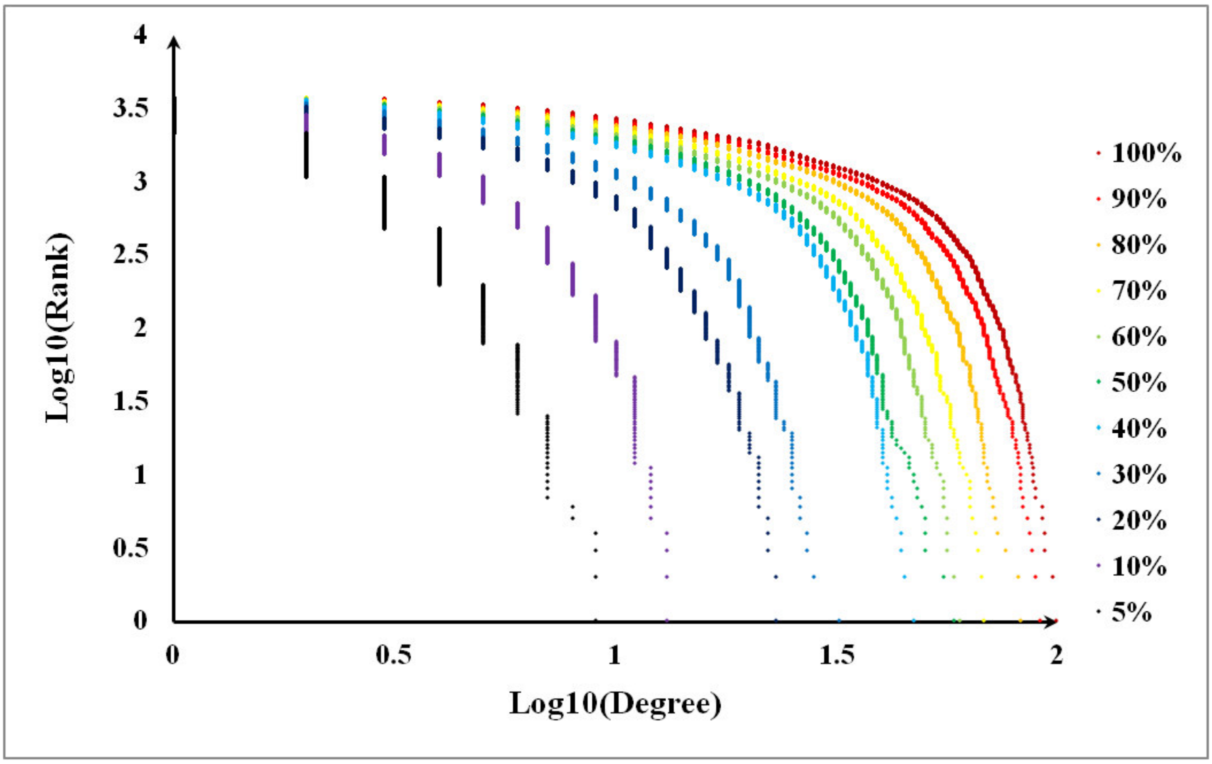


Figure 35: Log-Log plot of Rank of features sorted for decreasing connectivity (degree) over the degree itself. 100% relates to the full REMD set. 90% relates to 90% of the full REMD set (10% removed randomly) and so on. Linearity of the Log-Log plot indicates scale-freeness and a curved plot converges to Log-normality. A reduction of REMD number lets the degree distribution converge towards scale-freeness.

It is evident that the hypothesis is true, i.e. there are power topologies hidden in the network and it is imperative to devise methods of REMD reduction. At hand there are again two ways of REMD reduction: a knowledge based strategy and a data driven strategy.

4.3.1 The knowledge based strategy

The ultimate goal of the metabolome analysis – next to the definition of surrogate markers – is the definition of metabolic pathways, which are themselves mass difference networks. However, they are mass difference networks whose edges were validated over decades of experimentation. Conclusively, these validated pathways must have the core set of the metabolome embedded in their structure. It should be possible to extract the REMDs which differentiate a metabolic pathway from a random stoichiometric network.
In order to test this hypothesis, all available metabolic maps from KeGG as well as their

constituent metabolites were retrieved, all non-CHNOSP molecules and co-enzymes were omitted, and the list was concatenated and then networked. A second network with an additional rule was reconstructed afterwards: Edges were only allowed to be formed, if the nodes to be connected belonged to the same metabolic map. The first network was called holo-net and the second network was called inner-net.

The holo-net was taken to represent the frequency of each REMD over the entire population of KeGG metabolites. In analogy to gene set enrichment analysis (GSEA) [Subramanian, A., et al., 2005], the inner-net was taken to be a sample population. Consequently each REMD could be attributed with a frequency throughout the entire population and a frequency throughout the inner sample population that represented the validated pathways.

Consequently the Fisher exact test was applied in order to test whether an REMD was significantly enriched or associated to the inner-net or not. Again in analogy to GSEA, this procedure will from here on be called Mass Difference Enrichment Analysis (MDEA).

The results of this approach, knowledge based MDEA, are listed in the following table.

Table 4: p-values and z-scores for the inner-net; enriched REMDs

REMD	p-value	z-score
(de-)hydroxylation	0	21.76628
(de-)hydrogenation	0	17.67328
(de-)phosphorylation	0	17.46306
(de-)methylation	0	12.48104
hydrolysis/condensation	0	10.61447
deamination	0	7.742653
(de-)carboxylation	0	7.36
amino-function exchanged by hydroxyl function	0	7.11028
hydroxymethyl transfer	3.32E-09	6.095446
formyl transfer	1.1E-08	5.878452

The found results exhibited strong analogy with the functional REMD list as published by Tziotis [Tziotis, D., 2011] and they follow the major classification of Enzymes: Oxidoreductases, Transferases, Hydrolases, Lyases, and Ligases. (Isomerases are only detectable if the same mass occurs twice in one map, but since such redundancy was filtered out beforehand and because isomers cannot be distinguished by DI-MS, Isomerases were not considered to begin with).

Consequently the resulting REMDs were applied for network reconstruction of the smoker set.

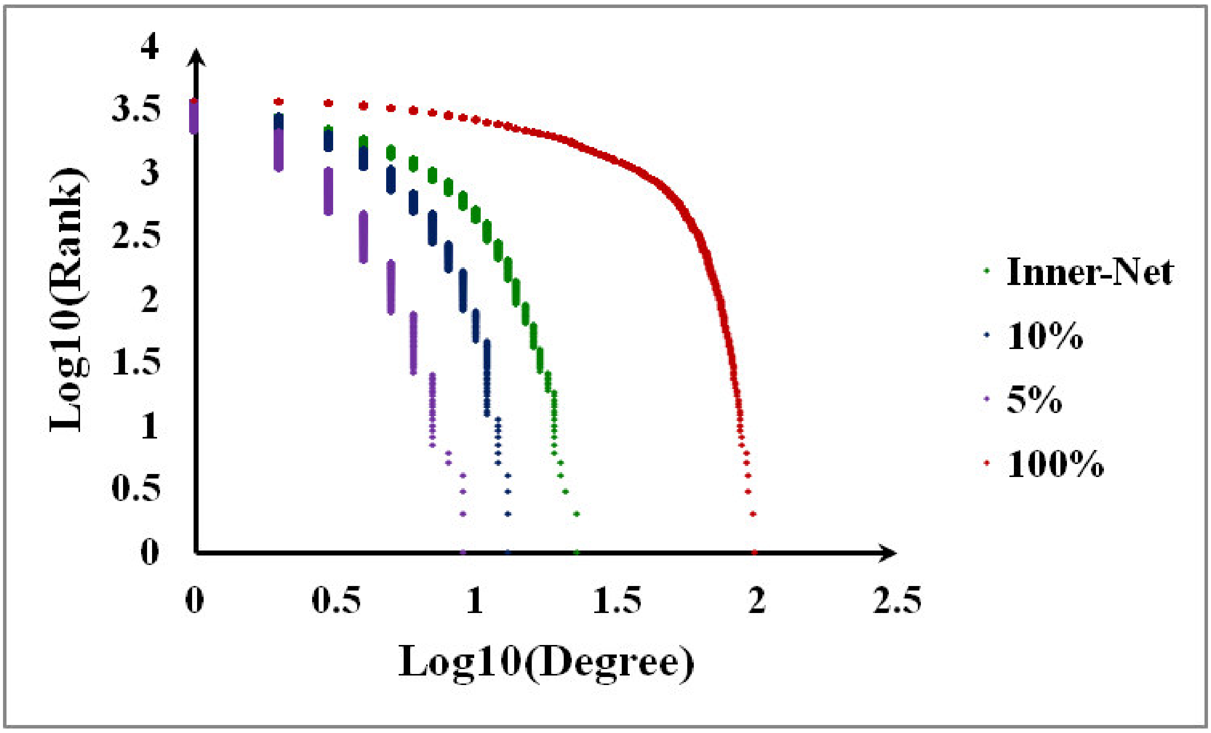


Figure 36: Log-Log plot of rank over degree derived from the Gauting network under use of the inner-net in comparison to the 5%, 10% and 100% of the full REMD set. The curvature of the inner-net is reduced in comparison to 10% and 100%.

The devised strategy enabled an improvement of network topology, but the approach in general is hypothesis driven. As discussed in chapter 1 and chapter 3, hypothesis driven research has the drawback of under-fitting the data, i.e. the danger of missing what is truly important is inherent. Additionally, it has to be asked, whether it is desirable to limit an REMD sets towards theoretical maps of metabolic pathway. As described in chapters 1 and 3, participants of one pathway cannot be guaranteed to occur in the same mass spectrum and reaction steps which are laid out in a stepwise manner in KeGG might in reality occur on enzyme-complexes without release of intermediates. This again nurtures the question as to whether metabolic pathway should be exclusively defined in a complete and exact theoretical context or whether it would not be beneficial to define more dynamic and data-driven pathways that reflect the empirical reality.

4.3.2 The data-driven strategy

We know that the reduction of REMDs improves network topology. Here, we complete the picture by means of data driven REMD reduction, which immediately allows for data interpretation.

MDEA of the Gauting results

The univariate analysis of the Gauting study revealed 169 discriminative features, of which 126 features showed a higher mean of normalized intensities in non-smokers. The multivariate analysis yielded two eigenvectors – E205 and E608 – whose directions were discriminative for smokers and non-smokers.

For E205 we extracted 82 features, which were conservative for its discriminative behavior. Positive eigenvector entries were indicative for feature co-occurrence in smokers and negative eigenvector entries were indicative for feature co-occurrence in non-smokers. Consequently, 49 features co-occurred in non-smokers and 33 features co-occurred in smokers. For E608 we extracted 46 features, which were conservative for discriminative behavior. In E608 positive entries were indicative for co-occurrence in non-smokers and negative entries were indicative for co-occurrence in smokers. There were 21 markers that co-occurred in non-smokers and 25 features co-occurred in smokers.

Up until this point, we did not assign any biochemical meaning to these features: neither did we assign metabolite names, nor did we assign metabolic pathways. We will now apply MDEA in order to improve network topology and in order to interpret, whether the respective enriched REMDs comply with the non-smoker phenotype and the smoker phenotype.

The REMDs, which were connected to non-smoker features and smoker features were tested for enrichment relative to the entire network. An REMD, which is enriched in non-smokers is not necessarily depleted in smokers.

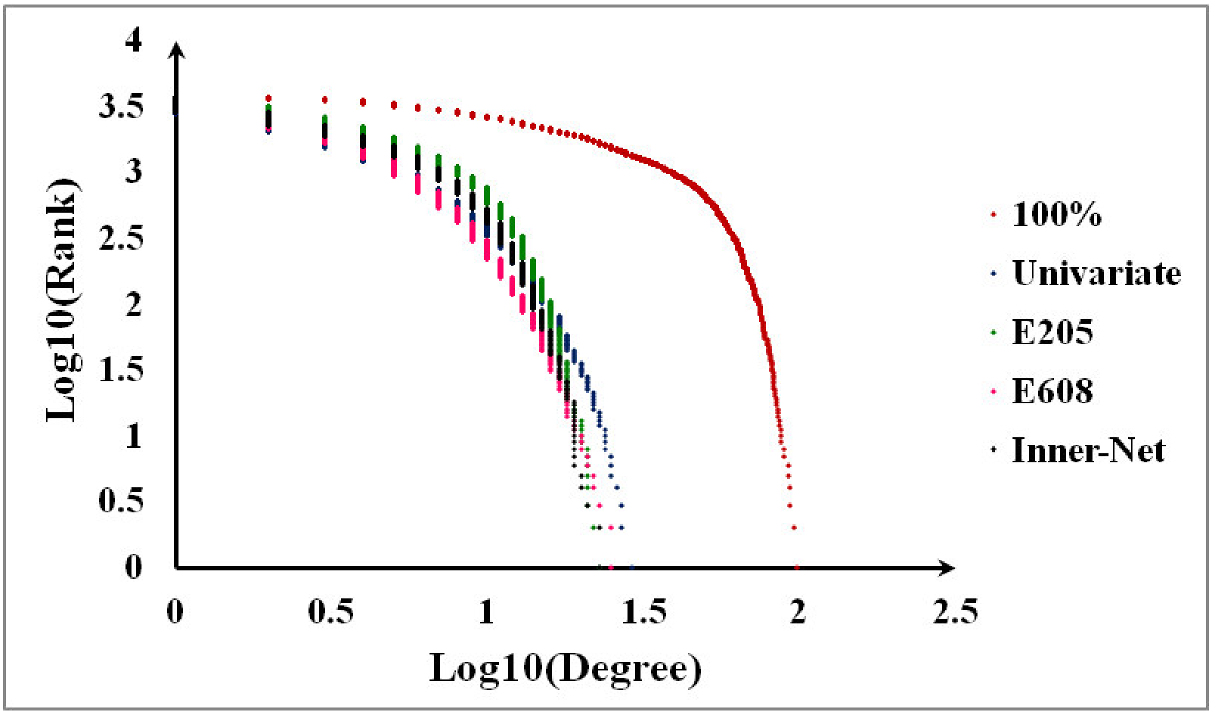


Figure 37: Log-Log plot of REMD reduction by means of data driven MDEA. All three marker scenarios show a less random characteristic than the network based on the entire REMD set. The univariate data suggests less curvature and a higher maximum degree than the inner-net, which indicates a higher degree of organization. E205 shows a stronger curvature than inner-net but has a lower maximum degree, which indicates more random behavior. E608 has a more flat curvature but a lower maximal degree than the inner-net, which indicates more organization.

The Log-Log plot of the data driven MDEA indicates an effective removal of random network connectivity, which is on the scale of the knowledge driven approach and more structured in the case of E608 and the univariate data. E205 indicates a strong random proportion.

Now let us use the REMD results for data interpretation. We first create a plot of univariate REMD results. Smoker and non-smoker reactions are plotted as bar charts with their magnitude being aligned to the enrichment z-scores.

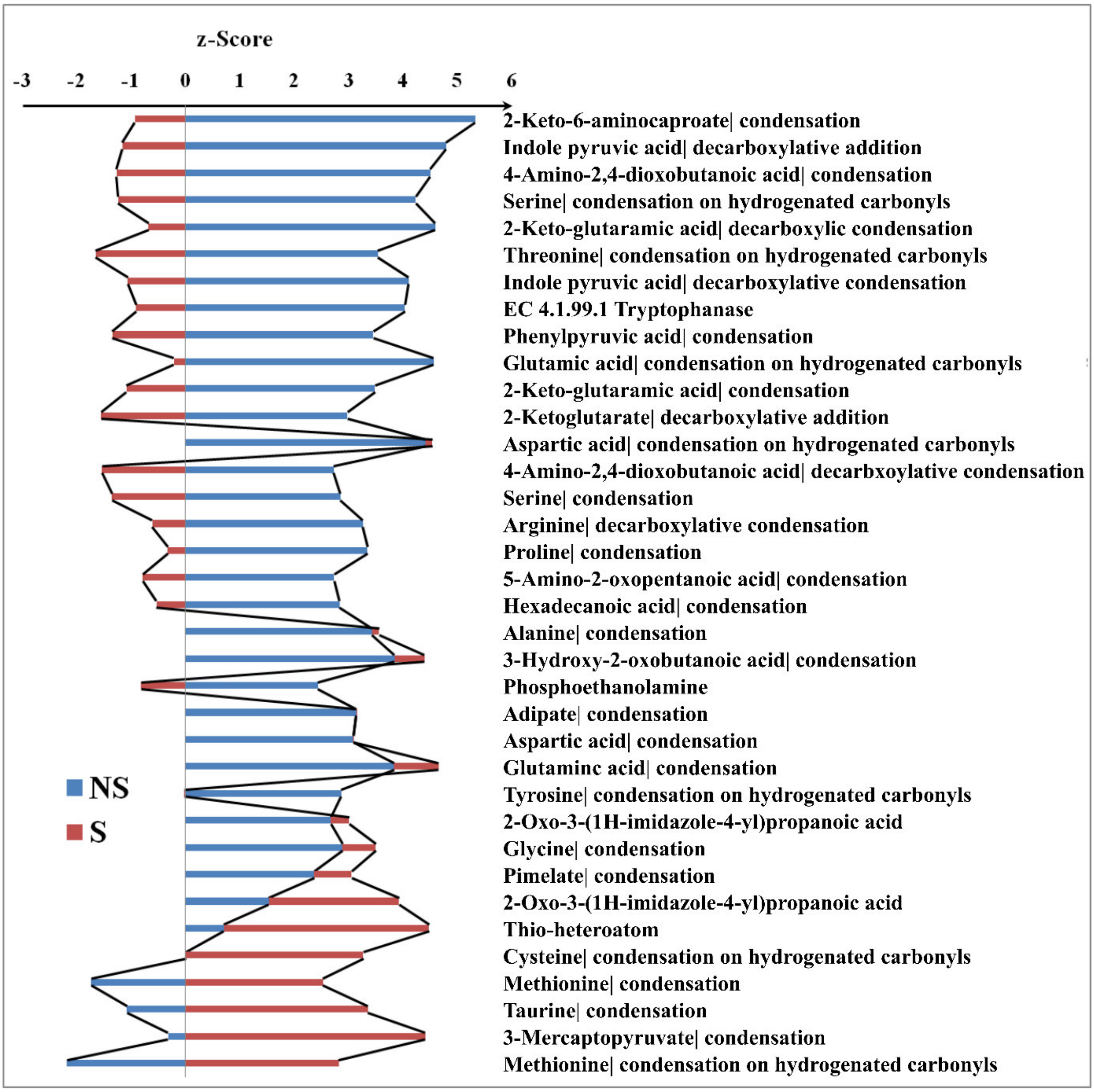


Figure 38: Vertical bar chart of z-scores for enrichment in univariate non-smoker (NS) and smoker (S) features and the respective REMDs.

The REMDs, which are the most associated to non-smokers, predominantly involve aromatic and basic amino acids and their corresponding keto-acids. Glutamate is considered to be the major energy source for macrophages in lung tissue and for lung tissue itself. Basic aminoacids may derive from blood plasma and may indicate the action of non-specific cation transporters, since they are mostly neutral at physiological pH. The REMD for EC 4.1.99.1 Tryptophanase is particularly interesting, since this enzyme is found in intestinal flora exclusively. This, however, is to be taken hypothetically as we have only detected a specific difference in elemental combination. REMDs of aspartic acid, glutamic acid, alanine, glycine and adipate are enriched both in non-smokers and smokers. These REMDs must therefore be important building blocks for EBC. Of special interest are the lowest six REMDs which are specifically associated to smoker markers. All these REMDs contain sulfur and are commonly involved in responses to oxidative stress such as methylation and hydrogenation. Cysteine, glutamic acid and glycine (all associated to smokers) are the building blocks of glutathione. Glutathione [$C_{10}H_{17}N_3O_6S$] itself was not detected but the given REMDs indicate significant involvement of sulfur compounds in response to cigarette smoking.

The enrichment analysis for E205 shows a different pattern.

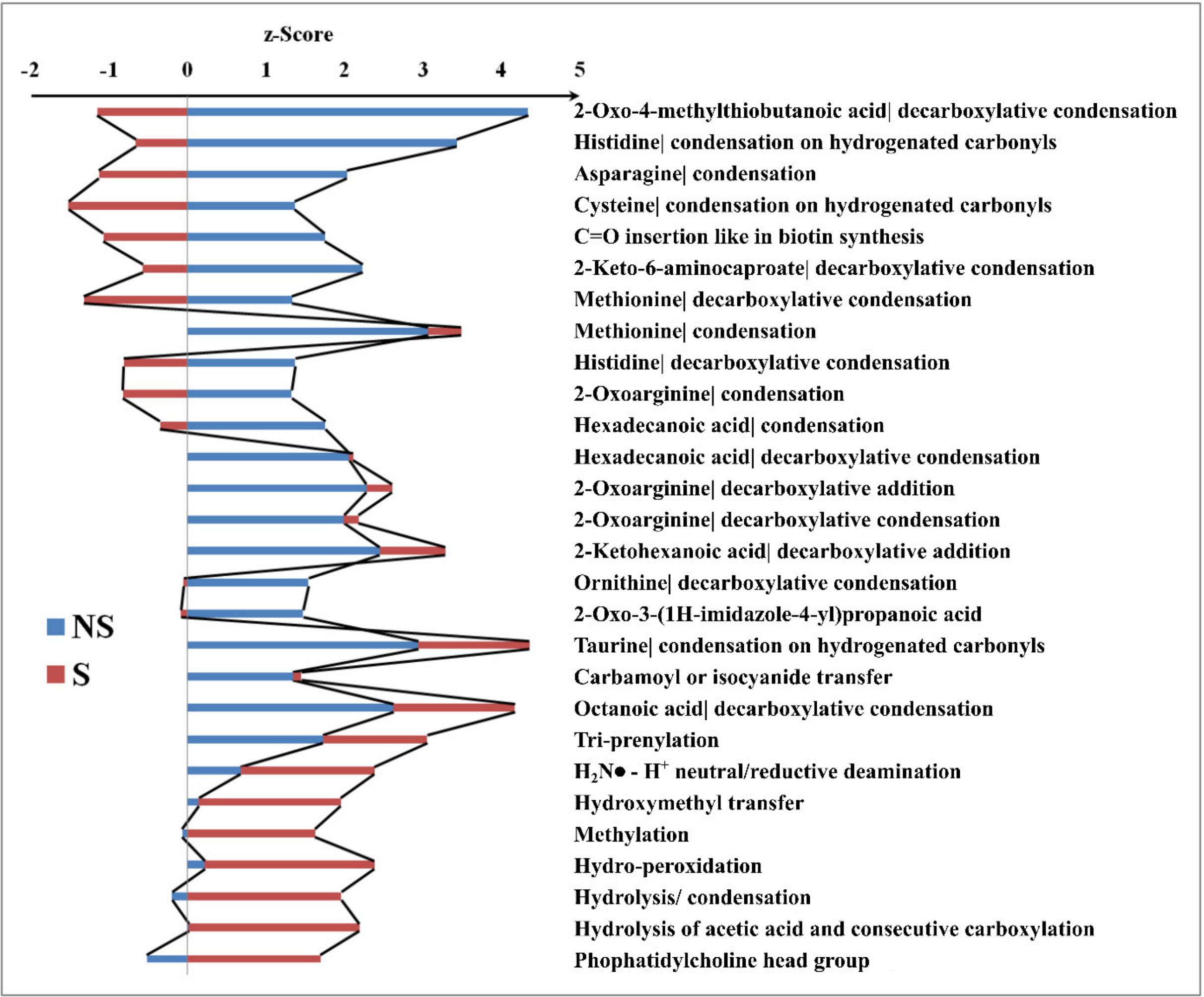


Figure 39: Vertical bar chart of z-scores for enrichment in non-smoker (NS) and smoker (S) features along E205 and the respective REMDs.

The pattern for non-smoker related compounds is enriched in sulfur containing REMDs, oxo-acids and some fatty acids. The more interesting pattern is again the lowermost part of the list, which relates to smoker features. Reductive deamination, hydroxymethyl transfer, methylation, hydro-peroxidation and hydrolysis alltogether are results of oxidative stress. Hydrolysis of acetic acid with consecutive carboxylation results in the simple addition of C; it may as well indicate double de-hydrogenation. The phosphatidylcholine head group REMD may relate to lysis of the epithelial membranes.

It is interesting that the features to which the enriched REMDs belong are not differentially regulated in the univariate sense but that they are ‘merely’ co-occurring in the respective phenotypic groups. This fact may be interpreted as follows: The chemical transformations, which pertain to these compounds, do not result in a stable end product. Instead, they are fastly being inter-converted due to excess of H_2O, H_2O_2 and radicals. This interpretation fits to the swiftly occurring redox reactions, which are emphasized by the MDEA smoker results and it fits to the stronger curvature of the log-log rank-degree plot of E205. Projecting this interpretation onto the non-smoker end of the given list, would indicate a generally high availability of sulfur amino acids, keto-acids, fatty acids and CO in the lung and it would indicate that these compounds are undergoing fast conjugation and disjugation with other compounds. These results give the impression of the lung being a very active chemical reactor, which is supported by the large surface area of the lung, as well as the basic activity of catalase and superoxide dismutase which is essential to aerobic metabolism.

The E608 features draw a more general picture of metabolism.

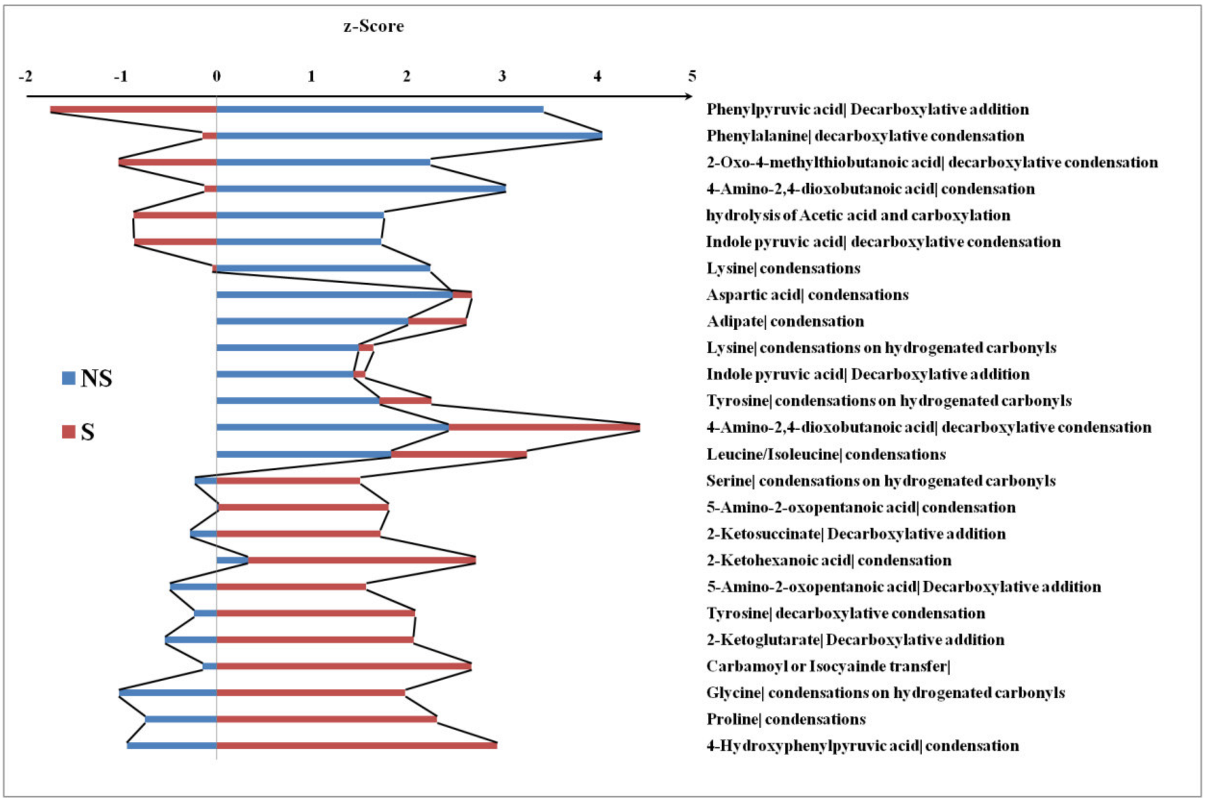


Figure 40: Vertical bar chart of z-scores for enrichment in non-smoker (NS) and smoker (S) features along E608 and the respective REMDs.

Along E608, aromatic amino acids and keto acids are conjugations that delineate the non-smokers phenotype. oxo-acids and cyanide conjugatioins are on the side of masses that co-occur in smokers.

Throuhgout all three marker classes, decarboxylative condensations are prominent relative to normal condensations. Univariate smoker markers show a different pattern. Those REMDs pertain to one-step condensations or to condensations after hydrogenation of a carbonyl group. This fact is interesting, because in this manuscript formal decarboxylative condensations were introduced in order to address pathways such as sphingosine synthesis, where serine reacts with palmitic-CoA under loss of CO_2 and H_2O. Such reactions depend on CoA or other thio-esters. The missingness of decarboxylative condensations in the univariate smoker-features indicates the absence of metabolism, which supports the common oxidative stress hypothesis.

Both strategies improved network topology in respect to the unfiltered network and they enable the data to choose which REMD environment is representative for the experimental context. The knowledge based approach does not allow for this freedom.

In an intersectional approach, we can now overlay a network, which was reconstructed based on the MDEA REMDs with correlation information between the features.

In theory the intersectional approach is bound to result in very small network components, since neighbors in metabolic pathways are bound to not-correlate if enzyme concentration stays constant. That is because enzymes are typically working under substrate saturation. The entire flux through a given pathway is managed by rate limiting enzymes, which means that they give a constant output and all enzymes downstream to the rate limiter are therefore independent unless regulation of the rate-limiting enzyme occurs. In addition, many metabolites have several roles or connections to other pathways, which alters their correlation to up-stream metabolites. There are attempts in literature which aim at the extraction of "the real" correlation between metabolic partners by means of partial correlation. If this was possible the network topology resulting from the intersectional approach would be close to equal to the MDEA network.

4.3.3 The intersectional approach

In order to create the intersectional Gauting network, we extracted all reactions, which were elements of all MDEA results. Then we calculated the Pearson correlations between the reaction pairs and made a cutoff at the 95th percentile of all resulting correlation coefficients. We omitted all reaction pairs that did not satisfy the respective correlation coefficient of $r^2 > 0.83$. The network was reduced from 16746 edges to 892 edges. The involved amount of nodes (features) was reduced from 3474 to 926 nodes. The number of previously defined differentiating masses was reduced from 286 to 54 features.

The resulting network was composed of 194 components of which only two components contained larger amounts of markers.

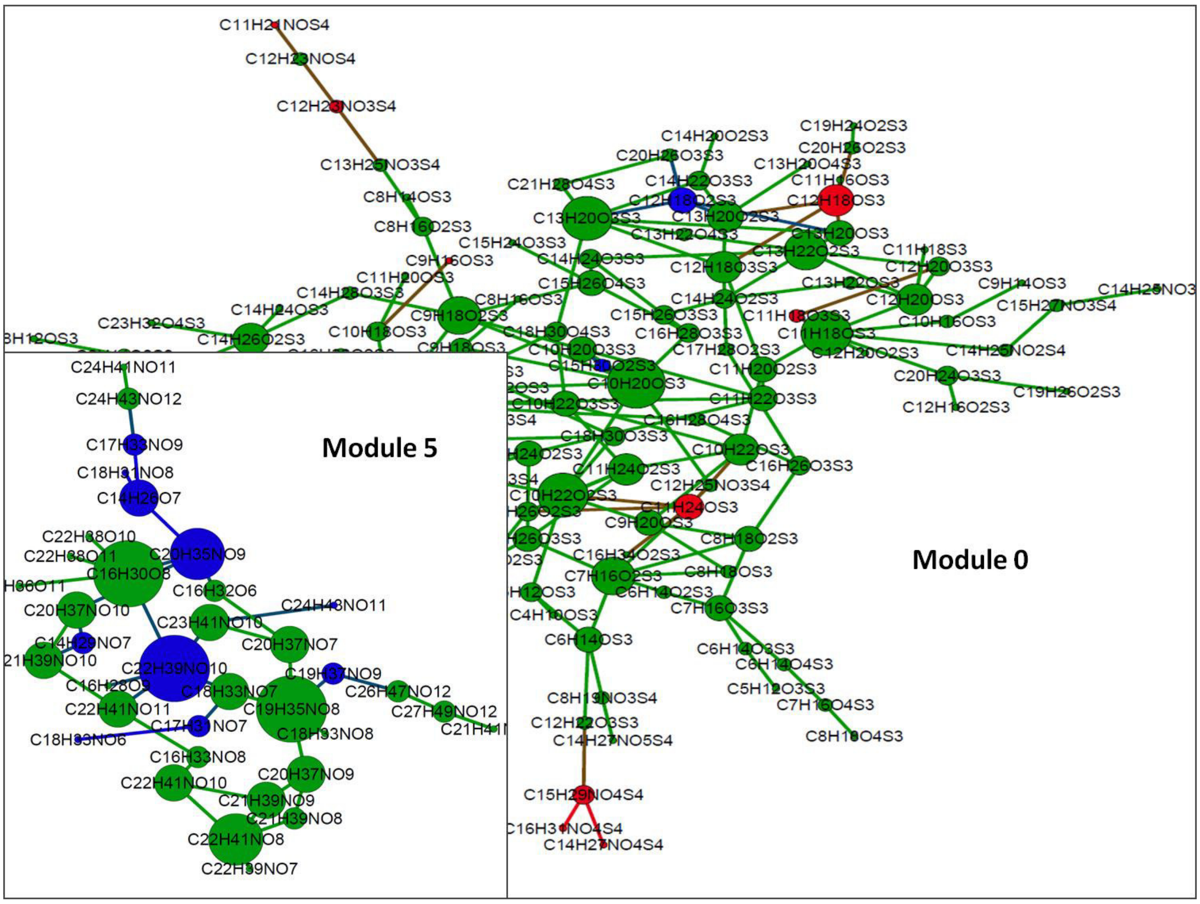


Figure 41: Intersectional Approach. Two modules with most markers. Module 5 relates to non-smoker markers. Nodes pertain to highly oxidized molecules with most REMDs relating to glutamic acid derivatives, pimelic acids and adipic acid. Module 0 is composed of sulfur rich nodes and pertains largely to smoker markers. An indicator for an oxidative stress response module.

Most compounds in both modules find no match in Chemspider. Module five REMDs relate to glutamic acid, pimelic acid and adipic acid. Some Chemspider hits indicate the involvement of hexoses. The REMD pattern of module 5 correlates to the univariate non-smoker enrichment results. The REMD pattern of module 0 is majorly composed of the REMDs which were found to be enriched for smokers both in the univariate context and E205.

In conclusion we can state, that random oxidative processes and high sulfur content are characteristic for features related to cigarette smoking. Non-smoker markers are characterized by the conversions, which are related to dicarboxylic acids and derivatives of glutamic acid and glutaric acids. In addition, respective markers are rich in oxygen and are conform to the finding of carnitines in EBC.

Another interesting fact is that non-smoker compounds are largely on the side of products in chemical reactions, while all other features are well balanced. Here, we assign the label 'product' to the larger mass in an REMD pair.

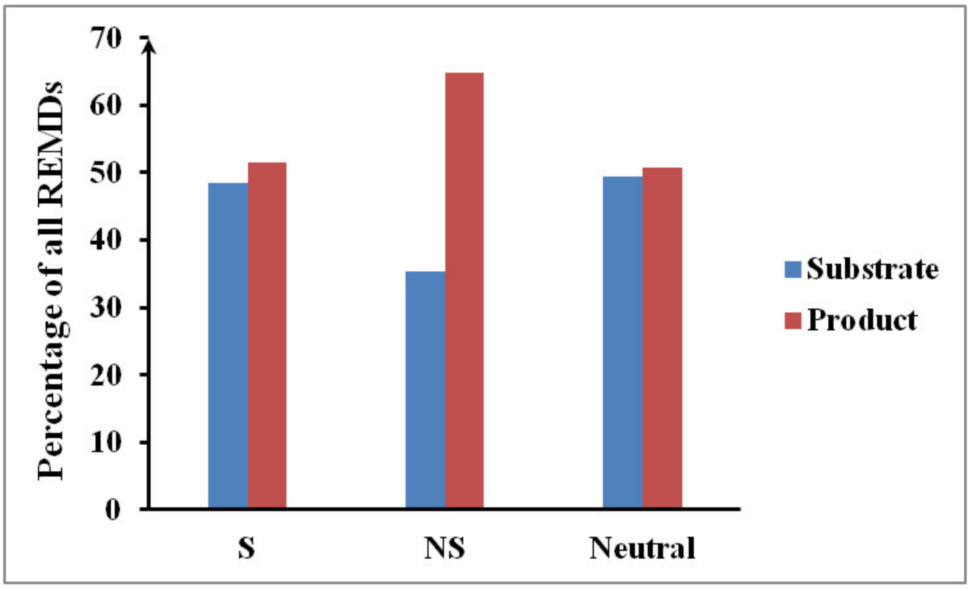


Figure 42: Balance between substrates and products in feature classes.

This finding can be interpreted in two ways: Non-smoker markers are either synthesized in EBC or they are constantly introduced into the alveolar lining fluid where they decay. Thermodynamically, 'de-novo synthesis' of non-smoker markers is improbable, which makes the decay scenario more probable. In consequence, we can hypothesize, that there is metabolite efflux either from pulmonary epithelium or from the blood system. This indicates that EBC may carry soluble markers from systemic metabolism.

On the other hand, non-smoker markers are either depleted or disorganized in smoking individuals, which underlines the effects of cigarette smoking. The balance between substrates and products among smoker surrogates as well as the high proportion of sulfur containing compounds indicates the prevalence of random, radical processes and a polarization of pulmonary metabolism away from primary metabolism and towards anti-oxidative action.

While it was known, that cigarette smoking and disease involve with oxidative stress, lung tissue is exposed to oxidative stress in healthy individuals as well (humans are aerobic organisms). While glutathione is involved into the normal response to oxidative stress, an extended use of other sulfur containing compounds was not pronounced in literature. A fact that was not presented above is that most common carnitines (15 hits) were annotated in EBC. The strong involvement of carnitines and compounds of similar composition in module 5 indicates that carnitines are strongly involved in normal lung metabolism. Real considerations regarding pulmonary energy metabolism and its relation to systemic energy metabolism were thus far only addressed in terms of glutamate. The present results contain an interesting set of markers for the investigation of pulmonary metabolism.

5. The HuMet Study: Links between EBC and Systemic Metabolism

The raw spectra of the HuMet study were calibrated against the rough Netcalc annotation of one positive mode HuMet EBC spectrum. The roughly annotated spectrum is not guaranteed to have been constituted of correct annotations only. The purpose of its use for calibration is to assist the visualization of the spectral error distribution, which is necessary to co-direct the error distributions of all 198 spectra. M/z spectra were then exported at S/N ≥ 3 and their alignment was intended to be performed at 1ppm error tolerance.

The attempt to align all 198 spectra into one matrix failed because the number of aligned variables by far exceeded the capacity of Microsoft Excel. It was therefore necessary to minimize the number of variables prior to alignment. The first filter was naturally based on de-isotoping.

Subsequently, we wrote an automatic filtering algorithm, that removes all m/z values whose absolute mass defect is chemically not realizable at charge z = 1 and given that only C, H, O, N, S, P and Na^+ are allowed. In order to provide a fast filter, which does not compare thousands of reference masses to thousands of reference values, we created a binary reference table of 1000 rows and 10 columns where each row refers to a down-rounded nominal mass and each column refers to a mass defect rounded down to the first digit. Based on a reference database of 18159 $[M+H^+]^+$ ions and 18159 $[M+Na^+]^+$ ions we filled each field of the reference map with "1" if it was host for at least three theoretical ions and we filled it with "0" otherwise.

The filtering then is performed by testing, whether the coordinate of an experimental m/z variable (nominal mass and down-rounded absolute mass defect are the coordinates) has the value "1" (valid) or "0" (invalid). Mass spectra containing 6000 to 10000 variables were filtered within one second per mass spectrum and yielded mass spectral read-outs of 40% to 50% of their original size. This filter ultimately enabled the unification of all 198 spectra into one matrix.

The unified matrix consisted of 62656 variables and 198 samples. Further data reduction was performed by means of Netcalc annotation. Netcalc annotation was performed in two stages. The data was first networked with a relative edge formation error of 0.1 ppm and a final error tolerance of 5 ppm.

Annotation was started at masses corresponding to sodium adducts of glucose and palmitic acid. Providing a large error tolerance increases the degrees of freedom; by experience, combinatorial algorithms and database matching techniques tend to randomly fill up the entire error space. Providing a large error space is implemented in order to validate whether the edge formation error is appropriate or not. If the error over m/z distribution is concise and centered after annotation (usually in the range of ± 0.5 ppm for more than 95% of the data), the EFE is not "oversampling" the data. If the error distribution 'leaks out' to the periphery, the EFE is too large and offers too many false edges.

The Netcalc algorithm used contained a filtering function, which deleted substrate and product formulae if a false annotation occurred. After setting these annotations to zero, the edge, which led to the false annotation, was getting marked. After an edge was marked five times, it was removed from the network. This algorithm extracts all sets of annotatable masses, converges to an average amount of annotations and finally oscillates around this number of annotations. Since each mass can have multiple isobaric annotations within a range of ± 0.5 ppm there are multiple scenarios of annotation which can lead to mass difference networks that contain almost no contradiction. The final annotation that Netcalc provides is not guaranteed to be the only solution. In order to pinpoint the most probable solutions, we printed the annotation results five times and extracted the most common annotations. Ultimately, we were able to mine an annotation set of 13124 features size.

Subsequent to annotation, all non-annotated m/z values were omitted, which is necessary in order to minimize co-linearity and in order to improve datamining.

Prior to datamining, it is necessary to investigate whether the uni-variate and multivariate data structures are biased. This investigation starts again with plotting the sample-wise sum of intensities over the sample-wise peak count. Figure 44 shows a dependency between both variables, which might be related to differential dilution of samples (matrix effect).

5.1 Investigation of data structure

We perform a first analysis of the data structure by using (multivariate) linear algebra. Similar to PCA, we want to use eigenvectors as representatives of data structure. The following steps need to be performed:

1) Create dataset D
    a. Raw dataset
    b. Normalize all rows on their Euclidean norm
    c. Normalize all columns on their Euclidean norm

2) Create the coincidence matrices of D by calculating **CM = D*D^T**

3) Calculate the eigenvectors of **CM**

4) Select the eigenvectors with the largest eigenvalues (e.g. usually the first three eigenvalues) and multiply them with all mass spectra

5) Plot the Eigenvectors against the frequency

6) Plot the resulting loadings over the samples

Similar to the analysis of the Gauting dataset, the Eigenvectors of the raw data reflect the typical power-structure of the data, for which reason plots versus frequency are not helpful. Normalization of the rows of the data matrix on their Euclidean norm results in more readable plots.

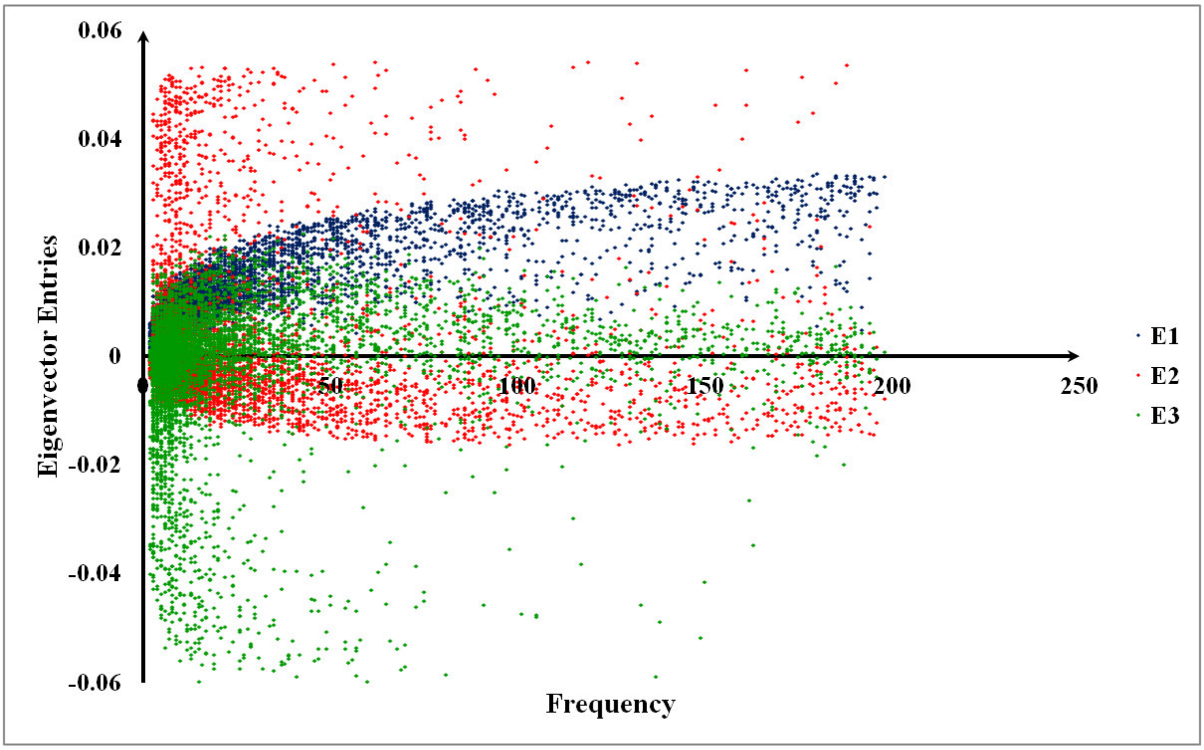


Figure 43: Plot of the first three eigenvector entries over the frequency of m/z values in the m/z-sample-intensity matrix. Large, positive eigenvector entries indicate a large degree of co-occurrence of an m/z with other m/z values throughout the samples. Large, negative entries indicate poor co-occurrence with other m/z values, i.e. singularities.

Visual inspection of figure 43: The largest eigenvector (E1) correlates with missingness, as expected. Other than in the case of the Gauting study, the second and third eigenvectors (E2 and E3) are unbalanced towards low frequencies. This is an indicator for the existence of feature groups that are rare, but that strongly coincide. In other words, the given coincidence structure indicates the presence of bias. Also the peak count versus intensity plot indicates that samples with larger peak sums contain a larger number of peaks. In addition, it is obvious that the peak count varies strongly and that the number of peaks per sample is lower than in the Gauting study. This effect is due to the less effective ionization provided by the used Apollo 2 ESI source. However, since measuring the HuMet study involved the measurement of more than 1000 samples both in positive and negative mode, an automated sampling system had to be used instead of the NanoMate robot. By experience, the NanoMate robot needs sample wise adjustment of ESI pressure and voltage in order to stabilize the pneumatically non-assisted spray.

Figure 44 shows the presence of quantitative outliers was well. Other than in the Gauting-case, additional evaluations of the data quality need to be performed; as normalizations and scaling on the original data are potentially introducing or even magnifying bias in the data.

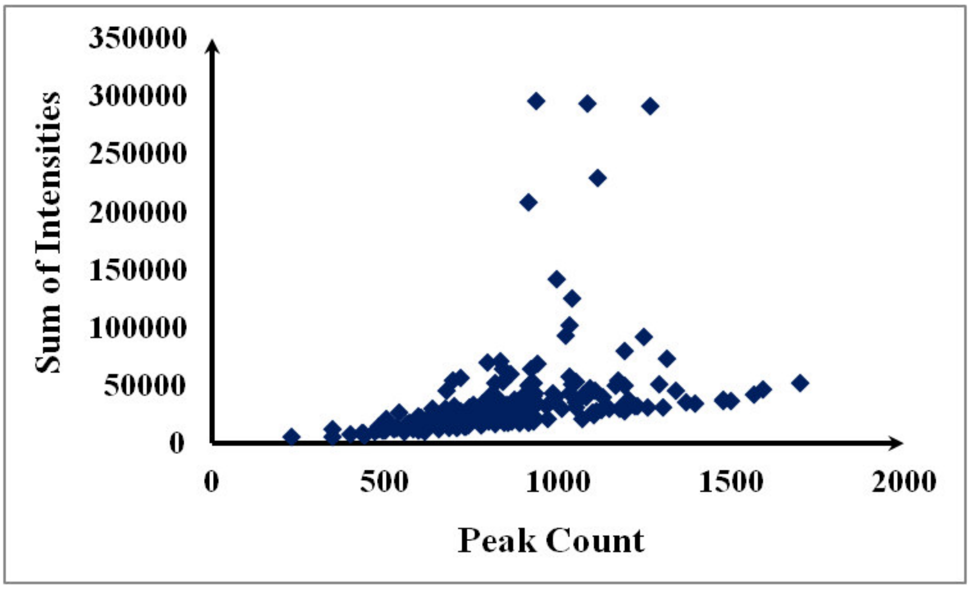


Figure 44: Sum of intensities over peak count.

Another intriguing point is a comparison of the eigenvector loadings against the peak count and sum of intensities. The respetive figures 45A) and 45B) show the eigenvector loadings and normalized peak sums of the raw spectra and normalized peak counts along the samples.

Interstingly, the peak counts are similar to the loadings of E1, which had a strong association to the frequency of m/z features over the samples. The Pearson correlation coefficient between both lists is 0.74. The loadings of E2 have a Pearson correlation coefficient of 0.72 with the normalized peak sums of the raw data. Let us recall, that E2 was strongly associated to bias caused by low frequency m/z features.

A fact that cannot be seen in this plot is, that the entries of large magnitudes associate with the specimen IDs. Since there was only one EBC sampling device, the specimens were sampled and stored batch-wise. The ICR-FT-MS measurements were performed batch-wise as well, since a random rearrangement of the samples may have lead to unwanted thawing of the samples.

The eigenvectors, which caused the loadings presented in figures 45A) and 45B) were generated on data whose rows were normalized. Therefore the magnitude structure along the mass list is cancelled out and the results mostly refer to the missingness structure and the quantitative structure along the samples.

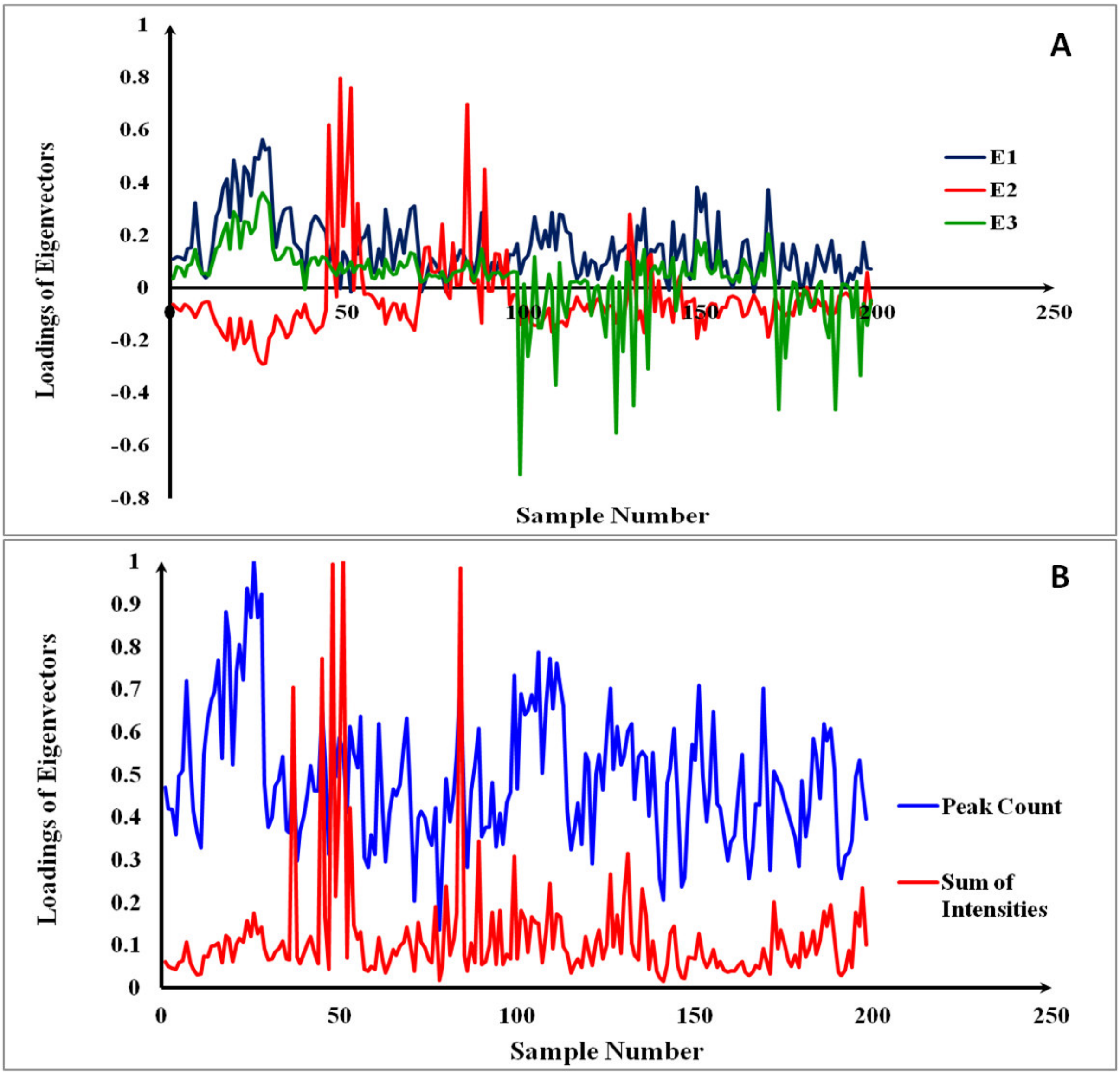


Figure 45: A) Plot of the loadings of E1, E2 and E3 along the sample list. B) Plot of peak count and sum of raw intensities, both normalized on the maximum norm, along the sample list.

Remark: Typical PCA analysis is said to require data that is centered, i.e. whose mean is set to zero. This step leads a) to centered PCA plots, which are esthetically more pleasing and b) to a co-intensity matrix, which is not only a co-occurrence matrix but also a covariance matrix. Eigenvectors, which are based on such a matrix always have a positive and a negative partition, where positive values stand for co-relation and negative values stand for co-relating feature pairs which co-relate but are anti-parallel to the main direction specified by the eigenvector (that is they are in anti-relation to the main direction). If such an eigenvector actually differentiates between sample groups, respective values can imply univariate up-regulation or down-regulation if and only if the data is complete. That means, if there are no missing values and complete case analysis is performed. However, this is an ideal scenario and does not comply with analytical reality. The lack of information upon this fact throughout literature sources leads to a systematic misinterpretation of multivariate results.

Non-centered PCA has its own advantages as we have seen. The first eigenvector of such an analysis is always positive and should relate to the missingness of a feature. Any other eigenvector should be independent of overall missingness structure but they can of course differentiate shifts of co-occurrence over the samples. A further advantage of non-centered PCA is that it is specific to co-occurrence. Centered PCA is centered about the sample mean, which is set to zero. That means, even if an m/z feature is specifically co-occurring in a group of samples, but it does not in another group of samples, its dot product may be negative due to the sign in the first group, but zero in the other group due to anti-occurrence. Non-centered PCA focuses the magnitude of an eigenvector entry on the amount of co-occurences and its sign towards co-or anti-directedness of the entry. *End of remark*.

In order to understand the magnitude by which different batches dominate the data structure and by which magnitude they therefore impede the analysis of metabolomics data, one has to create eigenvectors on the raw data.

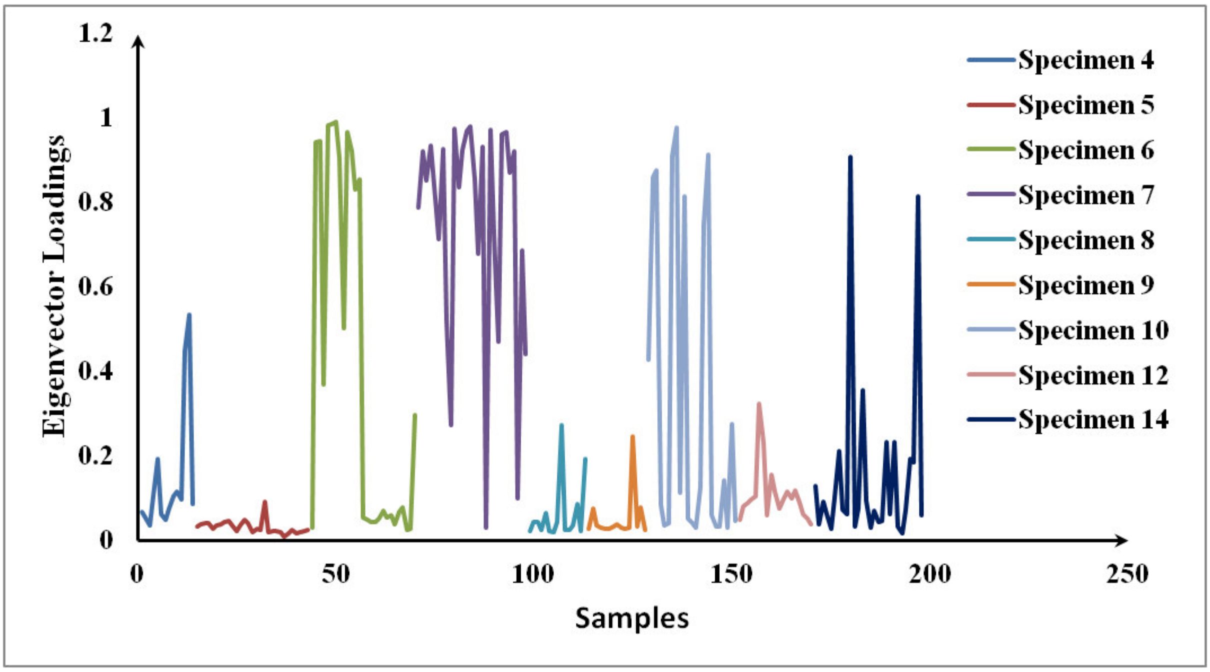


Figure 46: Eigenvector loadings of raw data. The large magnitude blocks relate to specimen 4, early specimen 6, specimen 7, specimen 10 and 14.

We can see that the magnitudes of coincidence, and of the involved intensities, are specific to the sample batch.

Now we know that the missingness structure is biased and relates to sample batches. We also know that magnitude plays a role and that the most varying variable is the peak count (except for some outliers pertaining to intensity).

In conclusion, further data cleaning has to focus on an investigation of batch-wise binary bias.

5.2 Elimination of binary bias

A mechanism by which binary bias can occur is varying sample composition (e.g. conductivity, dilution of the sample), which causes a varying ionization efficacy. In the previous subsection we found that there is a strong batch-wise bias in the data.

In order analyze the data, we need to remove the bias. If a dataset is homogenous, has no batch effects and no triplicate acquisition of samples, it is difficult to remove such bias. In that case, one would have to cluster the data in order to observe, whether strong clusters associate with low m/z feature frequency and whether the cluster is independent to the phenotypic classes of interest. If both conditions are true, an m/z value can assumed to be biased.

In the HuMet set we have a more comfortable situation, as we have proven strong batch effects. Removal of such binary bias can simply be performed by the following workflow:

1) count the frequency of non-zero elements per batch
2) determine the mean frequency and the relative standard deviation
3) calculate the z-score relative to the mean frequency for each batch
4) eliminate all features which contain a z-score > 1.96

The same can be performed using a Fisher exact test. We also know, that we are not interested into m/z features, which are binarily under-represented in all of the HuMet challenges (fastening [F], standard liquid diet [SLD], oral glucose tolerance test [OGTT], physical activity test [PAT] and oral lipid tolerance test [OLTT]). Assuming (axiomatically) that a non-zero frequency minor to 10% is not acceptable, all m/z features that have no acceptable frequency can be omitted.

The application of both filters yielded a dataset, which was reduced down to 2146 features (84% data reduction).

All steps performed up until this point pertained to data cleaning. The next step is normalization and we use the normalization algorithm, which was developed in chapter 3.

Other than the Gauting set, the HuMet set consists of a large amount of samples. Information that is not presented here indicates that fewer samples increase the number of correlations. Where the CIN based normalization cutoff for the dilution set (22 samples, $r^2 = 0.9$) conserved almost all features, the HuMet set had to be normalized at a cutoff of $r^2 = 0.5$ in order to conserve 85% of the given features [the omission of features during normalization occurs, when features have no correlation partner and can thus not be normalized].

In order to give an example of the efficacy of the CIN normalization approach, we have plotted the scaled intensites of the raw data, the Manhattan normalized data and the CIN normalized data (Figure 47).

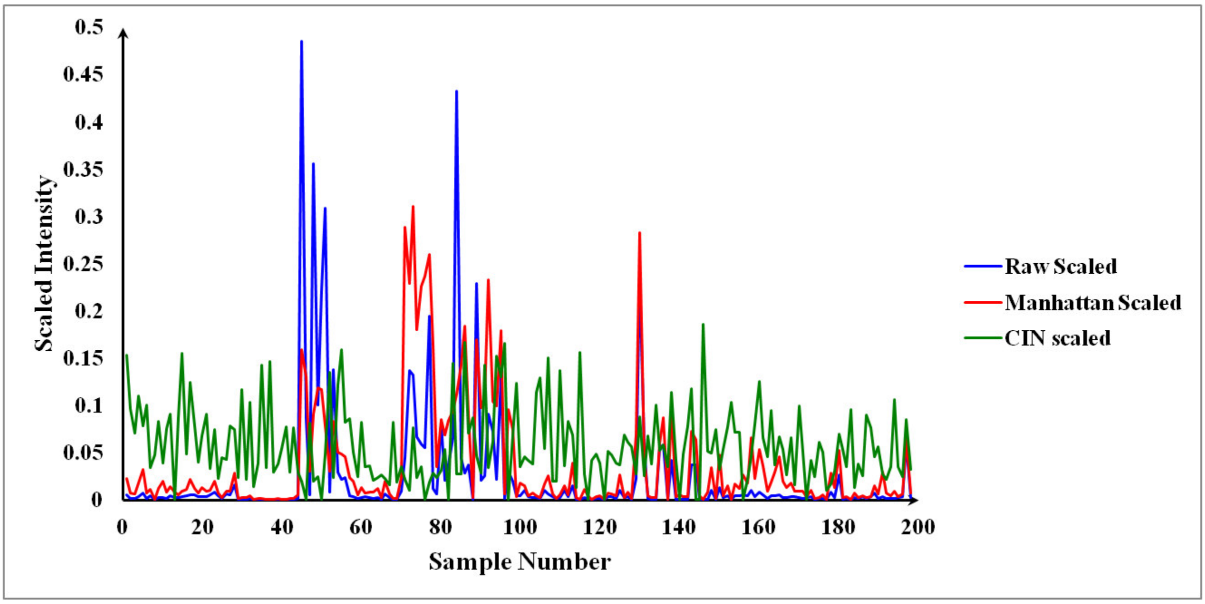


Figure 47: Normalization results over samples.

The CIN approach cancelled out the strong outliers. Manhattan normalization partially diminished outlierish peaks but it introduced bias as well.

5.2.1 Multivariate Analysis

We performed data analysis using eigenvector decomposition on non-centered data. Instead of creating a co-occurrence matrix over the features ($\mathbf{CM} = \mathbf{D}^*\mathbf{D}^T$), we created a co-occurrence matrix over the samples by applying $\mathbf{CM} = \mathbf{D}^{T}*\mathbf{D}$. The resulting Eigenvectors relate directly to the samples and they are almost identical to the loadings that can be yielded over $\mathbf{CM} = \mathbf{D}^*\mathbf{D}^T$. In the case of the Gauting set we did not present eigenvectors over the samples because the 25 eigenvectors yielded were not discriminative for smokers and non-smokers. In the Gauting set it was more effective to test the loadings into feature eigenvectors for sample discrimination.

In the case of the HuMet study we yielded 198 eigenvectors over the samples. Lacking a specific model for possible metabolite profiles, other than the ones published in Faseb, we inspected each eigenvector visually for any connection to the HuMet challenges. Since we investigate a time series with samples of up to nine different specimens, we smoothed the data using a running average over the samples, which were sorted for the challenges.

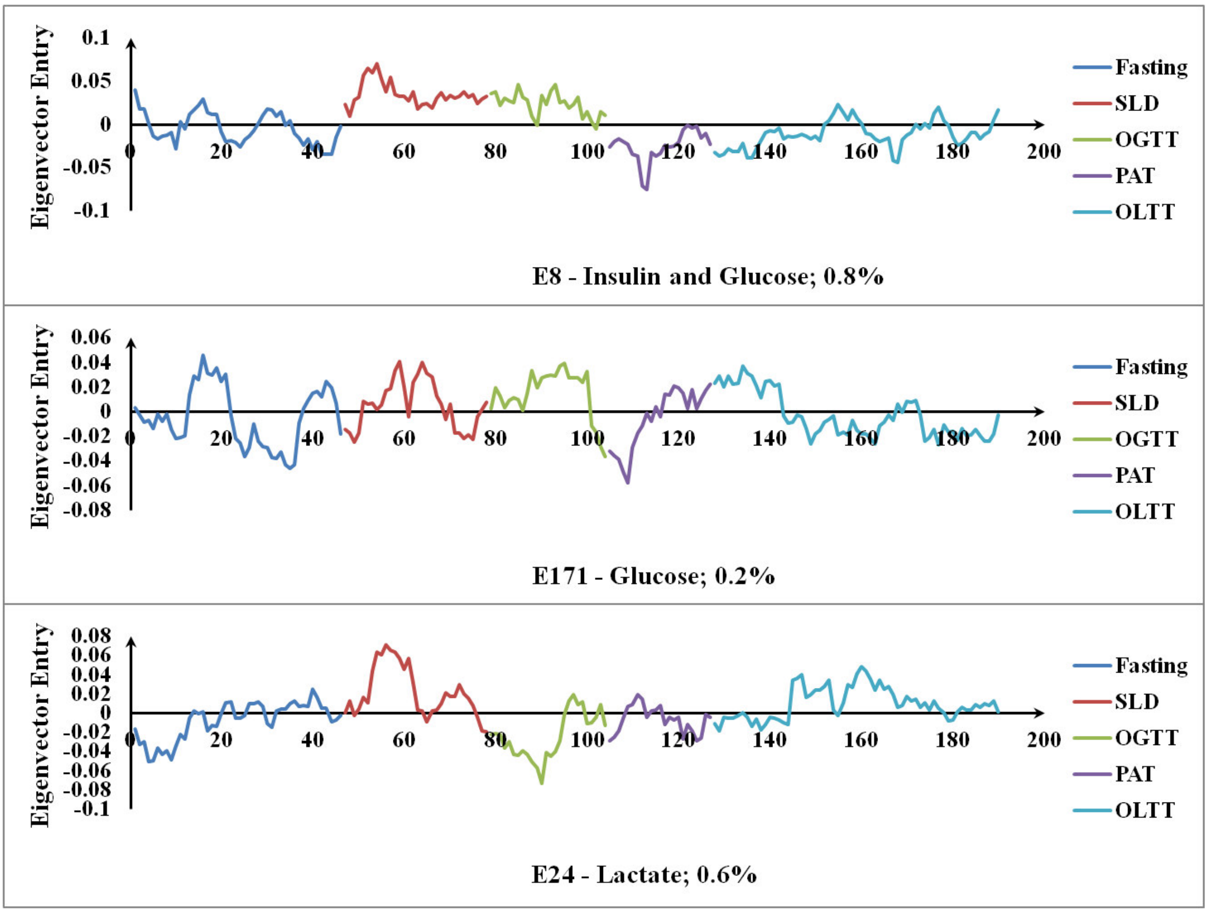


Figure 48: Three eigenvectors with the largest Pearson correlation towards the major clinical parameters determined by the HuMet consortium.

The displayed eigenvectors represent the most positive correlations with the insulin, glucose and lactate levels determined by the HuMet consortium. The eigenvector with the largest Pearson correlation to Insulin marked the maximum glucose correlation as well (r^2 = 0.51 and r^2 = 0.54). The second largest correlation with glucose marked eigenvector 171 (r^2 = 0.42). Most of the eigenvectors correlated with Lactate, with the maximum correlation being marked by eigenvector 24 and r^2 = 0.43. The given eigenvectors covered 1.6% of the data. The first eigenvector pertained to 21% of the data; the second eigenvector pertained to 2.1% of the data; the third eigenvector pertained to 1.6% of the data. The first three eigenvectors showed only weak differentiation of the different challenges. Most of the remaining 195 eigenvectors showed meaningful profiles like the plots of E4, E5, E6 and E10 (next figure).

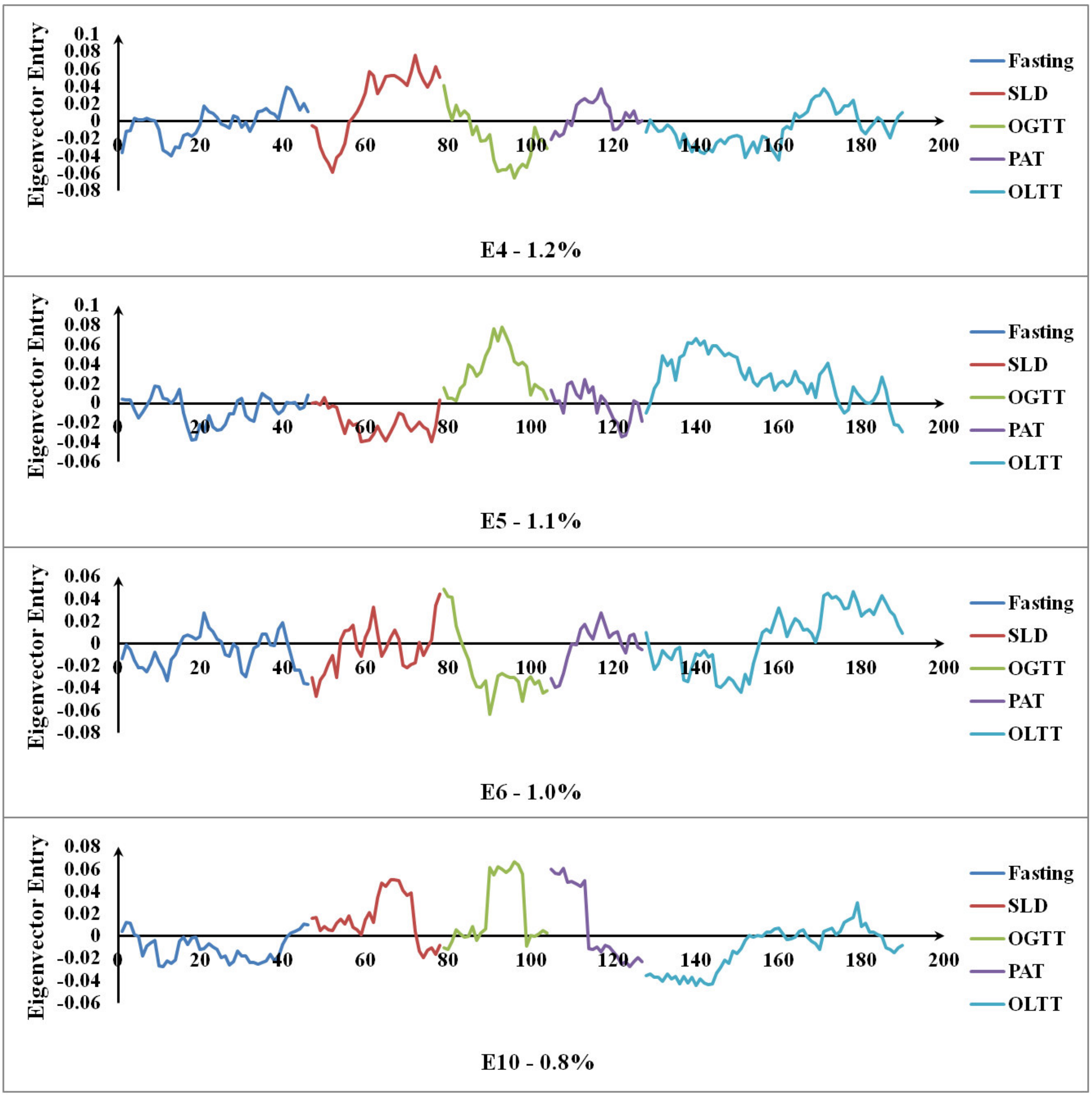


Figure 49: Eigenvector entries for the eigenvectors E4, E5, E6 and E10. Coloring of profiles according to HuMet challenge.

It is common to only regard to eigenvectors (or principal components) as being important, if they pertain to more than 90% of the data. However, this can only be realized if the largest amount of the data behaves accordingly. There are five challenges in the HuMet set and 'naively' assuming the possibility of two different states per challenge (1 → up-regulation, -1 → downregulation) would give a number of 2^5 (32) combinations. Adding statistical insignificance, i.e. orthogonality to the stimuli, there are 3^5 (243) combinations of statistical scenarios. Likewise, if all scenarios were equally probable, each eigenvector could only pertain to 0.41% of the data. It is clear, that the magnitude of an eigenvector is not of primary importance.

We can therefore interpret the eigenvectors as surrogate markers for a given scenario and we can state that it makes sense to first screen the eigenvectors, and to then extract the features which associate to them (load into them).

All eigenvectors listed are unit vectors. Their sum is necessarily the vector which halves all angles and shows the main direction of all main directions.

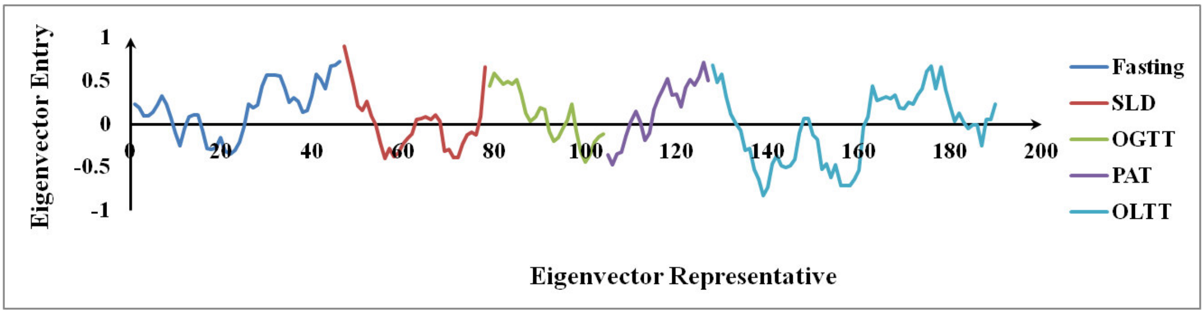


Figure 50: Eigenvector representative

The Pearson correlation coefficients of insulin, glucose and lactate to the eigenvector representative are $r^2_{\text{Insulin}} = -0.43$, $r^2_{\text{Glucose}} = -0.32$ and $r^2_{\text{Lactate}} = -0.46$.

Interestingly, the general direction of the HuMet dataset is exactly opposite to the most common clinical parameters.

Let us summarize the scenarios displayed by E4, E5, E6 and E10:

E4: Negative entries of E4 pertain to periprandial stages in SLD, OGTT and OLTT. Positive entries relate to the late phase of fasting, the post-prandial stages of SLD and OLTT and to the PAT challenge. They could therefore relate to the action of glucagon. Features, which are positively associated to this eigenvector seem to behave in the opposite sense to insulin.

E5: Postivie entries of E5 specifically react to OGTT and OLTT, while fasting, SLD and PAT do not induce a specific reaction. Compounds that relate to this eigenvector might specifically stem from the high concentration of hexoses and lipids in blood plasma.

E6: E6 behaves similar to E4, but the post-prandial phase in OLTT is pronounced more strongly.

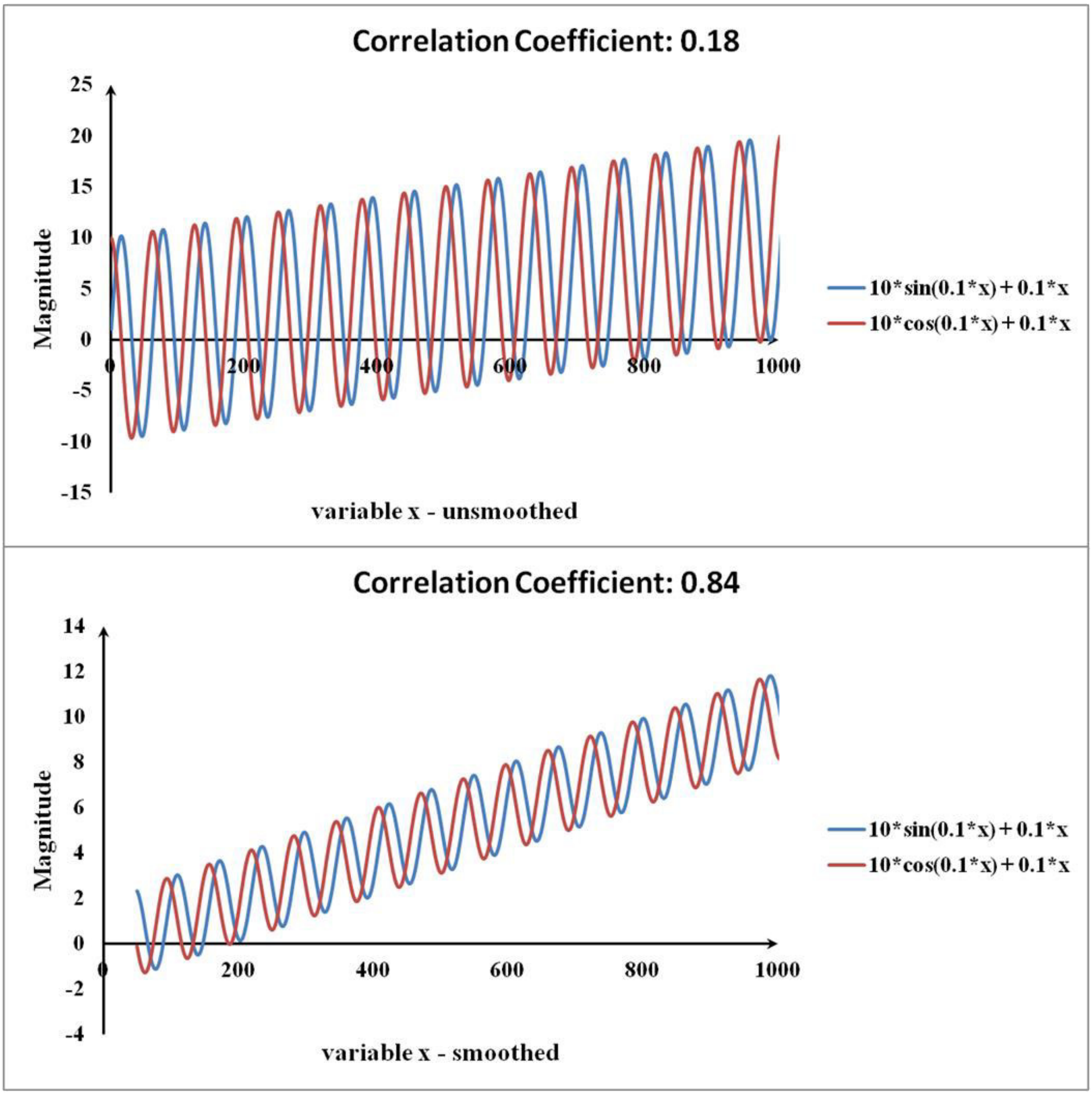
E10: E10 seems to be more specific to the post-parandial phases. We can hypothesize a strong involvement of glucagon, which would fit to the PAT profile as well. However, the OLTT response is weak.

Grouping of Eigenvectors

We had originally planned to perform experiments, which directly confront nutritional challenges against each other (e.g F vs. OGTT, F vs. OLTT, PAT vs. OGTT, PAT vs. OLTT) and to extract univariate and multivariate surrogate marker candidates in the same fashion as presented in chapter 4. However, we had to discover, that there were not enough samples for SLD, OGTT and PAT (at maximum 3 per specimen). We tried to investigate co-intensity matrices of such setups and discovered, that the connectivity of the resulting networks was dependent on missingness in any scenario that had less than 70 to 80 samples. For this reason we decided to extract marker candidates from the eigenvectors of the full dataset, which cover a sufficient amount of samples.

Consequently, we extracted the eigenvectors E4, E5, E6, E8, E10, E24 and E171 and grouped them together with all other eigenvectors that correlated to them (a supervised grouping of the eigenvectors or principal components).

Naturally, it is impossible to group eigenvectors, because all eigenvectors are orthogonal to each other.



Correlation Coefficient: 0.18

Correlation Coefficient: 0.84

Figure 51: Demonstration as to why smoothing enables the clustering of eigenvectors, which by definition are orthogonal. The linear equation y = 0.1*x was overlaid once with the trigonometric function y = 10*sin(0.1*x) and once with the function y = 10*cos(0.1*x). The pearson correlation coefficient of both functions naturally indicates orthogonality. Overlaying them with the linear equating gives a two functions whose oscillation center is monotonically increasing and that produce an r^2 of 0.18. We then performed average soothing. Each average encompassed 100 incident x variables. The result is a smoothed function with r^2 = 0.84.

This however, lies in the ‘microstructure’ of the eigenvectors, i.e. the specific co-behavior and anti-behavior of the singular samples. However, if the eigenvectors are smoothed by the moving average as we did for the eigenvector profiles above, we can cluster general directions.

We have correlated all smoothed eigenvectors with the smoothed E4, E5, E6, E8, E10, E24 and E171 and have chosen a correlation cutoff at the 95th percentile of all resulting correlation coefficients (r^2 = 0.28). We associated each eigenvector block (E-block), calculated their average direction and correlated each block representative to the smoothed data.

Since the intersection between E4 and E6 was large, the following groups and proportions of overall data resulted: E4|E6 (28.9%), E5 (7%), E8 (7.9%), E10 (3.3%), E24 (3.4%) and E171 (2%).

5.3 Association to smoothed E-Blocks

In order to evaluate, which features would be associated with which E-block, we first had to smooth the data as we had done it with the eigenvectors. After smoothing, we correlated each feature with each E-block and determined each correlation coefficient $r^2 > 0.37$ to be significant (95th percentile). The exceptionally large amount of database matches that we had acquired by performing MassTRIX annotation (37%) on the theoretical masses of the Netcalc output enabled the association of most E-blocks to specific metabolites.

For matters of space, we will only present and discuss the three largest E-blocks.

Block E4|E6

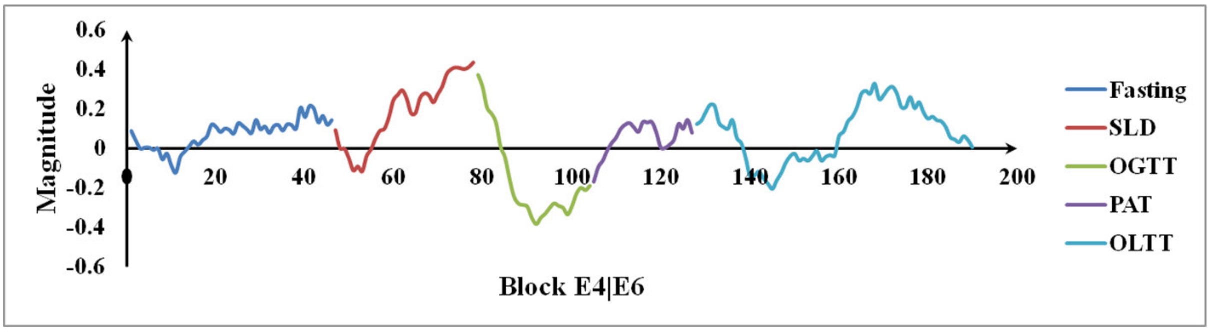


Figure 52: Representative magnitudes of the E4|E6 block. There is a strong post-prandial response in SLD and OLTT, as well as positive entries in PAT (the positive entries at the begininning of OLTT are an artefact from smoothing). In addition, the late fasting phase seems to be co-directed with the other positve phases.

Metabolites associated to the E4|E6 block encompass 10 out of 35 detected carnitines, while the respective E-block pertained to 196 of 1822 features (tiglylcarnitine, butenylcarnitine, 2-methylbutyroylcarnitine, O-propanoylcarnitine, pimelylcarnitine, 5-tetradecenoylcarnitine, decadienoylcarnitine, octenoylcarnitine, octanoylcarnitine, 3-hydroxy-5,8-tetradecadiencarnitine).

The expected proportion of carnitines covered by this block would have been four carnitines given a hypergeometric distribution.

Another prominent group of compounds was related to arachidonic acid derivatives and linoleic acid derivatives: 20-COOH-leukotriene B4, 12-keto-leukotriene B4 and (15S)-15-hydroxy-5,8,11-cis-13-trans-eicosatetraenoate as well as traumatic acid, 13(S)-

HPOT;(9Z,11E,14Z)-(13S)-13-hydroperoxyoctadeca-9,11,14-trienoic acid, 12-OPDA and (6Z,9Z,12Z)-octadecatrienoic acid. These compounds may very well relate to the PAT test, where increased oxidative stress occurs. In addition, the amino acids asparagine, tyrosine, phenylalanine as well as hippuric acid are related to the E4|E6 block. Asparagine REMDs were already reported in chapter 4 and the other amino acids were recently referenced to play a role in the development of diabetes type 2 [Suhre, K., et al., 2010; Würtz, P., et al., 2012; Cheng, S., et al., 2012; Wang, T.J., et al., 2011 and Huffmann, K.M., et al., 2009]. Since the OGTT profile of the E4|E6 block is clearly under-represented, there might be a connection between the E4|E6 block and glucose uptake.

Block E5

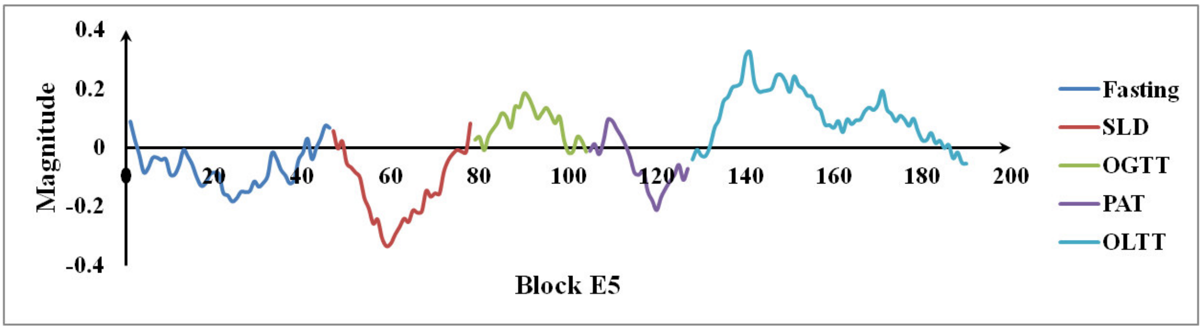


Figure 53: Representative magnitudes of the E5 block. Strong positive responses occur almost exclusively in OGTT and OLTT. Markers related to this block can therefore be assumed to be markers of high carbohydrate and lipid loadings in blood plasma.

The E5 block appears to relate to hyper-glycaemia and hyper-lipidaemia but without relation to Insulin. If Insulin action would have been involved, the SLD section would have had to show positive entries as well.

The annotations of the E5 block were almost exclusively composed of lipid variants. Acyl-lipid related markers encompass: heptadecanoyl carnitine, 2-(9Z-hexadecenoyl)-glycerol, 2-oxooctadecanoic acid, 2,6,8,12-tetramethyl-2,4-tridecadien-1-ol, hexadecanoic acid, 2,3-dihydroxycyclopentaneundecanoic acid, pentadecanoic acid, 12-hydroxydodecanoic acid, 2,4-decadienoic acid and 4,10-undecadiynal.

Sphingosine related marker candidates encompass: phytosphingosine, C17 sphinganine and hexadecasphinganine.

Other marker candidates were isomers of: oleoyl glycine and N-(3-oxooctanoyl)homoserine lactone.

The given marker candidates show signs of methylation (odd-numbered C count) and oxygenation (oxo acids).

No carbohydrate was found, for which reason we can assume that E5 exclusively targets lipid trafficking. The given compounds indicate that the lung is under constant oxidative stress, which is only natural. Since the HuMet study addressed the normal dynamic range of the metabolome, it is to be marked that EBC might not be the appropriate analytical matrix for the validation of hypotheses, which relate oxidative stress to diabetes mellitus or other pathologies belonging to the metabolic syndrome (oxidative stress is the 'working-horse' throughout almost any disease in the literature-landscape).

Block E8

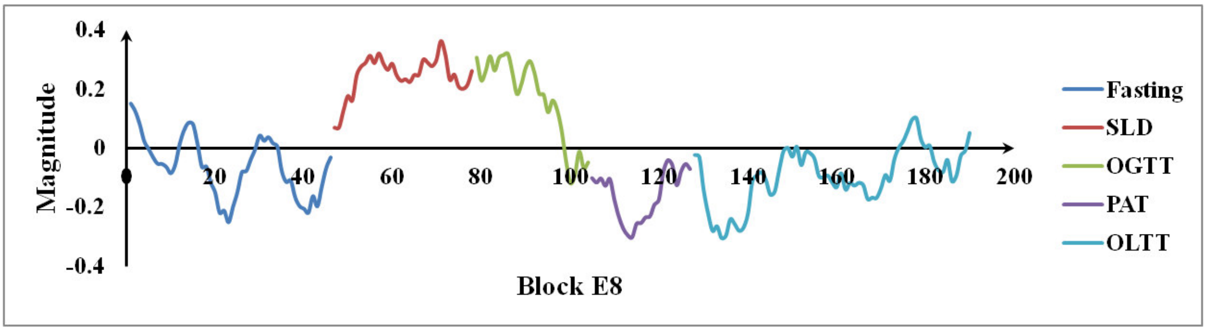


Figure 54: Representative magnitudes of the E8 block. Strong positive responses occur in the SLD and OGTT sections, exclusively. The fasting period, which covers 36 hours oscillates harmonically.

Interestingly, the E8 block oscillates harmonically in the fasting section. This observation may relate to the circadian rhythm and it had been a topic of past meetings of the HuMet consortium. The profile is specifically attenuated in the SLD and OGTT section, which indicates an involvement of amino acids and hexoses. The standard liquid diet used in the HuMet study was Fresubin®. According to www.DONG.de Fresubin contains 18.8% of carbohydrates, 5.6% of proteins, 5.8% of lipids and several vitamins and trace elements.

The annotaitons of the E8 block are well balanced throughout different compound classes. Amino acids and derivatives were represented by methylhistidine, L-serine and L-lysine-1,6-lactam and the carbohydrate glucose was found. We found the two steroids urocortisone and 7-dehydrodesmosterol as well as the terpenoid pentalenene. The sphingoid bases 1-deoxy-tetradecasphinganine and (4E,8E,10E-d18:3)sphingosine and the lipoamino acids tridecanoylglycine and pentadecanoylglycine as well as adenine were the major nitrogen containing findings.

The largest amount of annotations pertained to a versatile group of fatty acid derivatives ranging from medium chain length to C20: nonane-4,6-dione, linoleic acid, gamma-undecalactone, ethyl (R)-3-hydroxyhexanoate, pimelic acid, 5-hydroperoxy-7-[3,5-epidioxy-

2-(2-octenyl)-cyclopentyl]-6-heptenoic acid, 4,6,11-hexadecatrienal, 3-oxohexadecanoic acid, 3E,5E-tridecadienoic acid, 2-arachidonyl glycerol ether and 2-amino-9,10-epoxy-8-oxodecanoic acid.

This group of fatty acid derivatives underlines the oxidative stress baseline as several compounds indicate multiple oxidations. Regarding the presence of L-serine and multiple fatty acid derivatives (with cycles and lactones), it may be possible that spontaneous reactions produce the N-containing lipids of different chain length. Normally, the synthesis of sphingosine precursors requires the presence of pyridoxal phosphate and lipoyl-CoA but a presence of 2-oxo acids would enable the same mechanism (decarboxylation and condensation).

5.4 MDEA of E-Blocks

Based on the results of the E-block annotations, we expected to find the carnitine transformations, different transformations revolving around phenylalanine and tyrosine, as well as C_2H_4 units to be enriched for E4|E6. E5 was expected to show enrichment for C_2H_4 units. E8 was expected to be enriched in histidine, serine, lysine and all E-blocks were expected to be enriched in transformations that are typical for oxidative stress, i.e. methylations, peroxidations, nitrations or oxygenations, hydroxylations and dehydrations.

MDEA works by comparing the abundances of marker-associated reactions with the abundances of reactions throughout the entire population. The enrichment results for the E-blocks versus the entire population were poor.

E4|E6 was associated to arginine condensation on hydrogenated carbonyls, to deamination after dehydrogenation of hydroxyl functions, to decarboxylative condensation of 2-ketosuccinate and to C_2H_4 units (p-values between 0.001 and 0.02). E5 was associated with tryptophan condensation on hydrogenated carbonyls ($p = 0.015$) and E8 was associated with condensation of azelaic acid only ($p = 0.0002$).

Except for C_2H_4 units in E4|E6 the results did not match the expectations. In particular, no oxidative stress marker was associated with the groups. Initially, this is a surprising result, but considering that we compared the E-blocks versus the the entire population, we can interpret that oxidative stress and most other reactions simply belong to the normal reactomic spectrum associated to the lung.

Considering, that there is no differentiation between the entire population of reactions and the E-block reactions, we decided to unifiy the E-block reactions into a new reference population. This was done under the hypothesis that the effects of smaller differences between the groups were buffered by the entire population. Other than expected, there was no change in REMD patterns.

Considering that there are indeed compounds that differentiate between the different nutritional states, and considering that there are no changes in lung reaction patterns, the only logical deduction would be that the suggested surrogate marker candidates have their origin in blood plasma. We could assume an involvement of epithelial cell metabolism, but as we have observed above, E5 and E8 have patterns that imply insulin sensitivity. If at all, then E4|E6 implies Glucagon sensitivity. As we remember, we found strong polarization of REMD usage in the Gauting study, where the stimulus directly acted on the lung. But there is no polarization in the HuMet study. In consequence, we can state that we have successfully established a link between the EBC metabolome and systemic metabolism.

6 Summary

EBC is a 'blank page' in terms of metabolome analysis. The term 'deep metabotyping' was therefore chosen in order to underline the development of a workflow, which enables to extract a maximum of metabolically relevant information from EBC analyses. This goal encompasses feature annotation beyond the boundaries of metabolite databases, incorportation and neutralization of matrix effects as well as the extraction of biochemical context beyond database knowledge.

All these tasks are relevant for any analytical matrix, but they are of even greater importance once analytically 'under-described' matrices, such as EBC, are in the focus of metabolomics studies.

The interpretation of metabolomics data is always closely related to the origin of samples. EBC, whose origin is the airway lining fluid of lung epithelium, is complicated to interpret, because there are four different routes by which analytes may enter the airway lining fluid:

a) peri-epithelial transport from blood plasma to ALF
b) local metabolism of epithelial cells, pathogens or immune cells
c) immission of environmental chemicals
d) random (not genetically encoded) perturbation of a), b) and c) due to a highly oxidative environment; the presence of catalase, superoxide dismutase and cytochrome P-450 enzymes.

These different effects can normally be cancelled out or controlled in *in vitro* studies, or when other 'closed system' biofluids such as blood plasma or urine are analyzed. In traditional breath analysis, which focuses on volatile organic compounds, many aspects of the above four points can be cancelled out as well.

This is principally not possible in EBC analysis, and this fact complicates the extraction of systemic metabolism markers (as a matter of fact, systemic metabolism was never targeted outside of the HuMet study).

In this manuscript we have introduced Netcalc as a technique for m/z-feature annotation. We have introduced REMDs which are specifically designed in order to bypass the discontinuity of the metabolic chemical space. We have demonstrated that the safest methodology to perform calibration and database matching is to work on a fundament of Netcalc-annotated theoretical masses.

We have introduced routine key-points for the examination of mass spectral quality, both in the uni-variate sense and in the multi-variate sense.

We have established the hypothesis that the major proportion of variability of EBC metabolite abundances is due to dilution/matrix effect, and have shown that there are different classes of response to matrix effect. We have therefore deduced that it is not possible to adequately normalize data on one single dilution marker and have developed a co-intensity-network based technique for data normalization.

Based on the Gauting study we have demonstrated the nature of uni-variate and multi-variate marker candidates, which can be an important point in the definition of surrogate markers. We have demonstrated that mass difference enrichment analysis may aid in the assignment of biochemical/chemo-mechanistic context to lists of surrogate marker candidates (despite lacking database support).

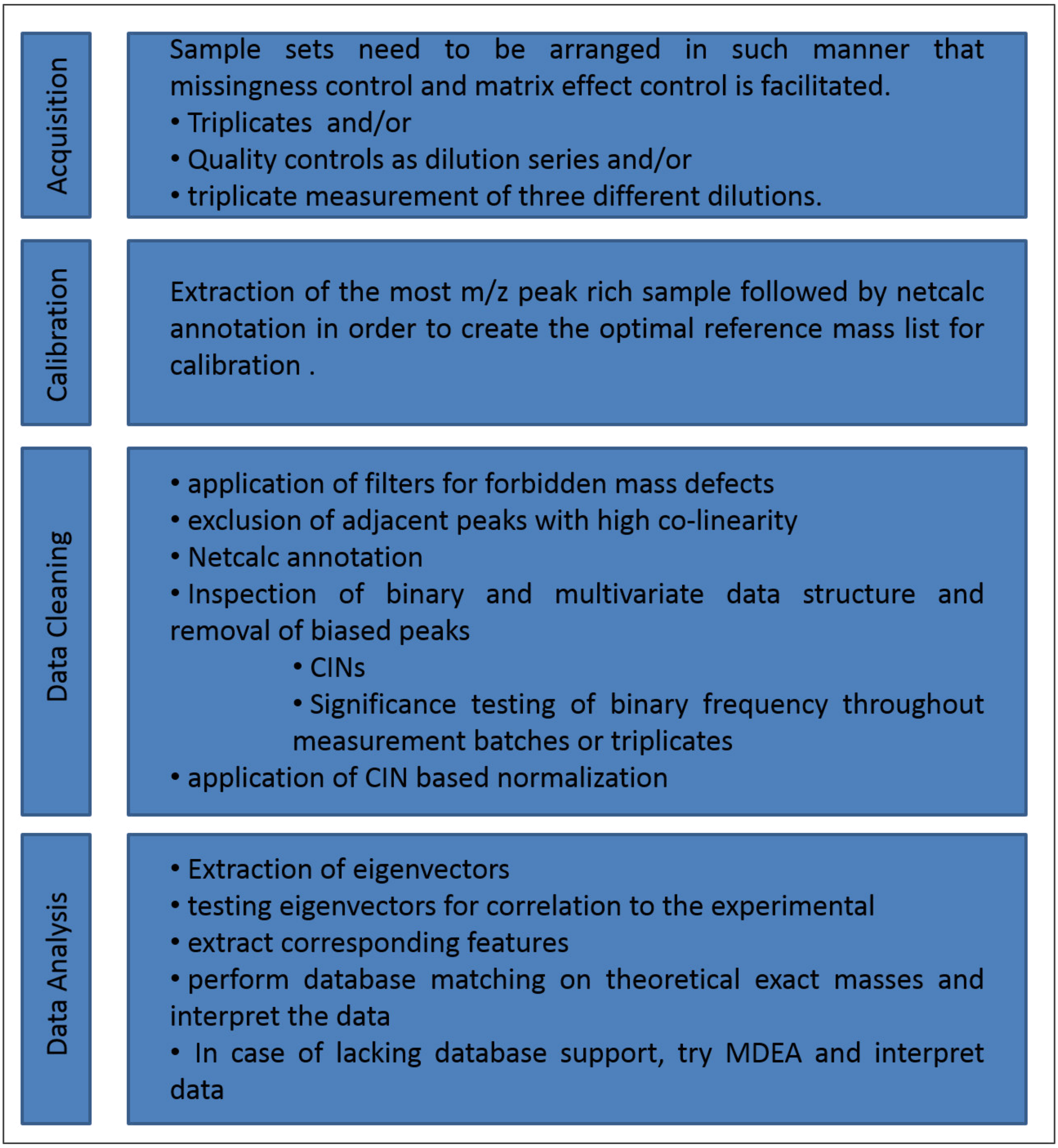
Using the HuMet study, we have applied all previously introduced techniques and emphasized the importance of missingness control. Finally, we were able to extract data eigenvectors (principal components) that were clearly associated to the different challenges of the HuMet study. Singular eigenvectors were specific to pre-, peri- and post-prandial phases of nutrition. Other than commonly assumed, the important information was not carried by the three first eigenvectors. Instead, the important information was spread throughout all eigenvectors, which is particularily due to the multi-challenge study design. The general direction of the data was shown to be anti-directed to insulin action and therefore co-directed to glucagon action. This assignment was possible particularily due to the application of moving average smoothing, which is a common technique in time series analysis.

In a next step we managed to cluster eigenvectors into E-blocks, which is originally not possible, since eigenvectors are orthogonal by definition. The moving average technique smoothed the effect of singular measurements and revealed the general data directions. Where the original eigenvectors covered a small part of the data only, the E-blocks covered significantly larger proportions and their profiles were more attenuated than the profiles of the single vectors (an example of a multivariate marker system).

A discovery at the side was that it is wrong to focus on the largest eigenvectors (principal components) only, because this is possible only if a major proportion of data variability is co-directed. We have shown that small eigenvectors are equally important in large study setups.

The rate of database-annotation of the Netcalc-pre-annotated HuMet set was extraordinarily high (37%; 10% are typical), which enabled us to assign specific metabolite groups to the E-blocks. MDEA of the E-blocks did not reveal interpretable patterns, which supports the hypothesis that E-block profiles are not of local (pulmonary) but of systemic origin. Found metabolite groups were in coherence with the findings of the original HuMet paper (Krug et al., 2012). Here the novelty laid not in the validation of the HuMet plasma proviles, but in the fact, that such profiles were found in exhaled breath condensate. In consequence we managed to suppert our hypothesis that EBC is a matrix suitable for the screening of systemic metabolism.

The workflow which results from our investigations is presented in figure 55.



Acquisition

Sample sets need to be arranged in such manner that missingness control and matrix effect control is facilitated.

- Triplicates and/or
- Quality controls as dilution series and/or
- triplicate measurement of three different dilutions.

Calibration

Extraction of the most m/z peak rich sample followed by netcalc annotation in order to create the optimal reference mass list for calibration .

Data Cleaning

- application of filters for forbidden mass defects
- exclusion of adjacent peaks with high co-linearity
- Netcalc annotation
- Inspection of binary and multivariate data structure and removal of biased peaks
    - CINs
    - Significance testing of binary frequency throughout measurement batches or triplicates
- application of CIN based normalization

Data Analysis

- Extraction of eigenvectors
- testing eigenvectors for correlation to the experimental
- extract corresponding features
- perform database matching on theoretical exact masses and interpret the data
- In case of lacking database support, try MDEA and interpret data

Figure 55: Data analytical workflow for deep metabotyping of EBC. The most important point is full control over, and removal of eventual biases.

It is of absolute importance to rather omit more data than too few data because binary bias cannot be corrected by any means of normalization. The dataset, which is used for multivariate analysis, has to show minimal dependency between the eigenvectors after the first eigenvector and feature frequency.

Marker candidates can be defined uni-variately and multi-variately. In non-full datasets it is not guaranteed for uni-variate manifestations of marker profiles to coincide with multivariate results. Instead, blocks of co-occurence may carry the information that was previously found in multivariate statistics.

6.1 Outlook

The present thesis exclusively pertains to the explorative stage of metabolomics, commonly termed ‘non-targeted’ metabolomics. In order to define surrogate markers, it is necessary to advance into the validatory or ‘targeted’ stage of metabolomics. This stage encompasses the development of an instrumental methodology of chemical analysis where the marker specific sensitivity is optimized and which enables the real identification of the respective surrogate marker candidates.

The development of such techniques can be based on the putative annotations of the pre-defined marker candidates. Here, we have found that different carnitines play a major role, especially in the E4|E6 marker block. This knowledge would indicate the use either of HILIC-UPLC-MS strategies or specific enrichment with cation exchangers. Such techniques would likewise enrich amino acids, which have moved into the field of interest.

Further marker classes that were found to be important especially in E5 were compounds whose stem-structure seemed to be unsaturated fatty acids which were then modified by different pulmonary oxidative mechanisms. Such metabolites would be more optimally analyzed by different reversed phase LC techniques.

Given the impact of dilution and matrix effect, future DI-ICR-FT-MS studies should either be measured in triplicates or there should be a dilution-quality control. Normal quality controls have a constant setup because the original purpose of their use is to validate the reproducibility of a workflow. However, if dilution and matrix effect prevails, there is no use of a quality control that always shows the same result because it does not address the actual bias. A dilution series of a quality control could in turn address this matter (as we have shown in chapter 3). Triplicates provide control over missingness but they do not provide control over the matrix effect. In consequence future setups should contain an intelligent constellation of quality controls and dilution series. It is as well imaginable to combine the triplicate measurement with a dilution row.

Further methodological developments pertain to the sampling strategy. The EBC device used for this thesis was the EcoScreen 2 sampler which uses Teflon bags for condensate accumulation. These Teflon bags bear sources of contamination, since the EBC is not conveniently to be withdrawn from them. In addition, we have found that too low condensation temperatures only increase the amount of H_2O vapour co-condensation, which ultimately adds up to the analytical problems pertaining to EBC analysis. We assume a sampling technique, which maximizes the collection of exhaled aerosolic droplets and minimizes the co-condensation of H_2O vapor to be more suited for non-invasive breath sampling.

It is to be emphasized that the analyte spectrum of EBC analysis is fundamentally different from that of classical breath analysis, which focuses on VOCs. EBC analyses reported in literature rarely report information that exceeds the measurement of H_2O_2 concentration or pH. We have demonstrated, that the analyst has a far wider spectrum of analytes at hand. The facts that EBC sampling is non-invasive and that it is possible to screen systemic information with EBC at hand should encourage its use for clinical studies.

Further developments in the field of data analysis should focus on raising awareness towards the mathematical mechanisms underlying these techniques, since interpretability of results should be in focus instead of the maximization of the complexity of applied techniques. Softwares provided by vendors of mass spectrometers encourage the use of thechniques without understanding their mechanisms.

In this thesis we could show that even perturbations pertaining to the normal range of metabolism can be addressed by EBC analysis. There are good prospects for the detection of diseases because they mark abnormal states of metabolic regulation. Special interest should be placed at the investigation of trans-epithelial transport processes, since the transport of non-volatile compounds through airway epithelium is clearly underaddressed. In addition, our findings suggest the investigation of the catalytic capacity of the lung. Beyond the datasets presented in this manuscript, random perturbation of EBC constitutents seems to be a common trait to EBC dynamics.

7 Supplementary Information

Methods:

Gauting:

EBC was sampled using the EcoScreen 2 device (Jäger, Germany). Sampling was performed at -20°C for 10 minutes.

Samples were stored at -80°C.

Samples were thawd on ice.

Sample dilution: EBC: MeOH = 1:2.

NanoMate - ESI: voltage and pressure were adjusted so as to deliver a constant ionization current of 10 to 20 nA.

Samples were acquired in negative mode over 1000 Scans at 1MW.

HuMet:

EBC was sampled using the EcoScreen 2 device (Jäger, Germany). Sampling was performed at -20°C for 10 minutes.

Samples were stored at -80°C.

Samples were thawd on ice.

Sample dilution: EBC: MeOH = 1:10.

Apollo 2 - ESI: samples were ionized in positive mode at 4500 V

Samples were acquired in positive mode, 400 Scans at 2 MW.

Table 5: Univariate smoker markers

Experimental m/z	Theoretical m/z	Formula	Ion Type	Erro [ppm]	p-value	Cyclomatic Number
445.2108572	445.211152	$C_{21}H_{29}N_6O_3P$	H^+	-0.662	0.0018	12
368.1742256	368.174618	$C_{16}H_{33}NO_2S_3$	H^+	-1.066	0.0087	7
340.143196	340.143318	$C_{14}H_{29}NO_2S_3$	H^+	-0.359	0.0106	7
415.098495	415.098339	$C_{16}H_{18}N_2O_{11}$	H^+	0.376	0.0111	9
569.3010392	569.301215	$C_{28}H_{48}N_4O_2S_3$	H^+	-0.309	0.0147	13
525.1401016	525.140224	$C_{21}H_{36}N_2O_3S_5$	H^+	-0.233	0.0152	15
341.0904795	341.090667	$C_9H_{23}N_2O_6SP$	Na^+	-0.550	0.0172	3
296.0801319	296.079851	$C_{10}H_{17}NO_7S$	H^+	0.949	0.0185	5

455.107196	455.107477	C16H19N6O8P	H+	-0.617	0.0206	12
312.05767	312.057874	C11H21NOS4	H+	-0.654	0.0216	10
386.1204814	386.120275	C16H23N3O4S2	H+	0.534	0.0223	11
402.089332	402.089569	C14H27NO4S4	H+	-0.589	0.0223	10
436.204796	436.204573	C21H36NO3SP	Na+	0.511	0.0238	8
345.0948311	345.094513	C11H21O10P	H+	0.922	0.0286	3
387.205652	387.206053	C17H30N4O4S	H+	-1.036	0.0296	7
375.2102891	375.210075	C21H30N2O2S	H+	0.571	0.0297	10
430.1206022	430.120869	C16H31NO4S4	H+	-0.620	0.0297	10
436.298315	436.298192	C18H42N7OSP	H+	0.282	0.0297	5
342.0856795	342.085331	C11H19NO9S	H+	1.019	0.0301	5
511.331094	511.331253	C26H46N4O4S	H+	-0.311	0.0309	8
448.1100071	448.110304	C19H29NO3S4	H+	-0.662	0.0388	14
358.0630636	358.063354	C12H23NO3S4	H+	-0.811	0.0397	10
375.2085091	375.208557	C12H30N4O9	H+	-0.128	0.0397	0
385.1169692	385.117164	C15H28O5S3	H+	-0.506	0.0402	8
277.1110875	277.111289	C13H24S3	H+	-0.727	0.0406	8
436.3032455	436.303345	C23H43NO5	Na+	-0.228	0.0410	3
314.061926	314.061893	C11H11N3O8	H+	0.105	0.0411	8
386.112025	386.112413	C14H27NO5S3	H+	-1.005	0.0417	8
531.320655	531.321083	C25H46N4O6S	H+	-0.806	0.0423	7
291.0751144	291.075387	C11H11N6O2P	H+	-0.936	0.0429	11
350.200128	350.200438	C17H35NS3	H+	-0.885	0.0434	7
375.194317	375.19482	C17H30N2O5S	H+	-1.341	0.0442	6
247.242015	247.242026	C18H30	H+	-0.044	0.0442	4
487.1858071	487.185713	C20H30N4O8S	H+	0.193	0.0442	10
377.2120825	377.212087	C20H34O3S	Na+	-0.012	0.0456	6
375.2142468	375.214196	C20H32O5	Na+	0.136	0.0461	5
415.10516	415.105351	C20H19N2O6P	H+	-0.460	0.0472	14
436.2958009	436.295853	C27H37N3O2	H+	-0.119	0.0472	11
419.2430825	419.24312	C19H31N8OP	H+	-0.089	0.0474	10
416.1048304	416.105219	C15H29NO4S4	H+	-0.934	0.0475	10
228.0179	228.018118	C6H13NO2S3	H+	-0.956	0.0478	7

237.043404	237.043604	C9H16OS3	H+	-0.844	0.0480	8
295.04886	295.049084	C11H18O3S3	H+	-0.759	0.0499	9

Table 6: Univariate non-smoker markers

ExpMass	TheoMass	Formula	Ion Type	Error [ppm]	p-value	Cyclomatic Number
331.1723177	331.172445	C12H22N6O5	H+	-0.384	0.0009	5
304.1514525	304.151649	C13H17N7O2	H+	-0.646	0.0012	9
304.1682947	304.168558	C15H22N5P	H+	-0.866	0.0019	9
304.1502971	304.150313	C12H21N3O6	H+	-0.052	0.0034	4
432.2228764	432.22281	C20H33NO9	H+	0.154	0.0053	5
345.1545629	345.154396	C16H24O8	H+	0.483	0.0057	5
304.1666778	304.166534	C11H27N3O3S	Na+	0.473	0.0065	2
306.1911467	306.191115	C14H27NO6	H+	0.103	0.0067	2
263.151557	263.151906	C11H23N2O3P	H+	-1.326	0.0071	3
321.2337614	321.23349	C17H34N2S	Na+	0.845	0.0072	4
299.05351	299.053515	C10H19O4S2P	H+	-0.017	0.0076	7
268.0895675	268.089543	C11H20NOSP	Na+	0.091	0.0079	6
304.1909377	304.19072	C18H25NO3	H+	0.716	0.0088	7
304.1718044	304.171929	C12H26N5SP	H+	-0.410	0.0089	6
463.2649214	463.265009	C21H38N2O9	H+	-0.189	0.0094	4
463.2592546	463.25935	C21H42N4OS3	H+	-0.206	0.0095	9
459.1456857	459.145701	C18H35O5S3P	H+	-0.033	0.0096	9
346.1860271	346.18603	C16H27NO7	H+	-0.008	0.0102	4
304.1730819	304.17306	C12H27NO6	Na+	0.072	0.0112	0
380.206765	380.206765	C20H29NO6	H+	0.000	0.0115	7
331.179245	331.179457	C16H23N6P	H+	-0.640	0.0119	10
427.1766363	427.176097	C19H32O7S	Na+	1.262	0.0126	6
511.1378629	511.138095	C22H26N2O10S	H+	-0.454	0.0126	13
360.2015983	360.20168	C17H29NO7	H+	-0.227	0.0127	4
321.15455	321.154396	C14H24O8	H+	0.480	0.0128	3
298.1861867	298.18603	C12H27NO7	H+	0.525	0.0134	0
304.1707819	304.170556	C16H27NOS	Na+	0.743	0.0139	6

238.1649217	238.1649	C10H23NO5	H+	0.091	0.0150	0
321.2204215	321.22064	C15H32N2O3S	H+	-0.680	0.0151	3
331.1864108	331.186364	C15H26N2O6	H+	0.141	0.0158	4
380.2643325	380.26428	C18H37NO7	H+	0.138	0.0163	1
390.2122117	390.212245	C18H31NO8	H+	-0.085	0.0166	4
396.2228568	396.22281	C17H33NO9	H+	0.118	0.0169	2
348.1499042	348.15004	C11H25NO11	H+	-0.390	0.0172	0
511.1310827	511.1311	C21H34O6S4	H+	-0.034	0.0173	13
331.1679871	331.168224	C17H23N4OP	H+	-0.715	0.0186	10
321.2324643	321.232524	C22H28N2	H+	-0.186	0.0186	10
296.183076	296.18323	C14H27NO4	Na+	-0.520	0.0186	2
690.3487596	690.348405	C36H51NO12	H+	0.514	0.0190	12
265.1028833	265.103029	C9H16N2O7	H+	-0.549	0.0192	3
439.265715	439.265876	C20H42N2O4S2	H+	-0.367	0.0197	5
668.2377279	668.238008	C32H45NO8S3	H+	-0.419	0.0202	17
393.2272714	393.227166	C22H32O6	H+	0.268	0.0207	7
171.1379728	171.137956	C10H18O2	H+	0.098	0.0207	2
543.2914867	543.291224	C26H42N2O10	H+	0.483	0.0207	7
414.1649317	414.165113	C23H28NO2SP	H+	-0.438	0.0210	14
368.2348638	368.234907	C20H34NO3P	H+	-0.117	0.0212	6
374.238333	374.238946	C18H36N3OSP	H+	-1.638	0.0215	6
353.1081938	353.108327	C16H21N2O3SP	H+	-0.377	0.0221	11
304.1825735	304.182477	C18H26NOP	H+	0.317	0.0231	8
304.1801656	304.180172	C12H25N5O2S	H+	-0.021	0.0231	5
438.2532113	438.253337	C21H35N5O3S	H+	-0.287	0.0235	9
314.0907246	314.090416	C10H19NO8S	H+	0.982	0.0241	4
317.19577	317.195866	C16H28O6	H+	-0.303	0.0247	3
481.2106552	481.210782	C18H39N2O7SP	Na+	-0.263	0.0250	4
357.2485391	357.248782	C19H37N2SP	H+	-0.680	0.0251	6
360.2379882	360.238065	C18H33NO6	H+	-0.213	0.0253	3
362.2172804	362.21733	C17H31NO7	H+	-0.137	0.0258	3
368.233031	368.233253	C22H29N3O2	H+	-0.603	0.0258	10
434.2384555	434.23846	C20H35NO9	H+	-0.010	0.0261	4

557.2724667	557.272692	C28H40N6O2S2	H+	-0.404	0.0270	16
352.196501	352.196595	C15H29NO8	H+	-0.267	0.0275	2
455.155024	455.154791	C21H26O11	H+	0.512	0.0280	9
321.2245092	321.224662	C20H32OS	H+	-0.476	0.0288	7
394.3163463	394.316315	C20H43NO6	H+	0.079	0.0291	0
246.01071	246.010924	C6H15NOS4	H+	-0.870	0.0295	8
348.1561829	348.156265	C18H25N3S2	H+	-0.236	0.0301	12
336.0856088	336.085999	C11H17N3O7S	H+	-1.161	0.0311	7
414.1517829	414.151575	C18H27N3O4S2	H+	0.502	0.0312	11
173.026706	173.026692	C7H8O3S	H+	0.081	0.0315	6
348.1483	348.148403	C16H29NOS3	H+	-0.296	0.0317	9
371.10478	371.104973	C13H19N6O3SP	H+	-0.520	0.0318	11
348.1970391	348.196771	C18H31NO2S	Na+	0.770	0.0321	6
305.1726579	305.172331	C16H26O4	Na+	1.071	0.0324	4
471.1442269	471.144047	C21H30N2O4S3	H+	0.382	0.0325	14
516.2802695	516.280325	C25H41NO10	H+	-0.107	0.0326	6
561.3310514	561.331648	C26H48N4O7S	H+	-1.063	0.0326	7
569.2422417	569.242082	C25H43N2O7SP	Na+	0.280	0.0327	9
165.0757804	165.075751	C6H12O5	H+	0.178	0.0327	1
525.2993488	525.299285	C27H44N2O6S	H+	0.121	0.0330	9
330.0682896	330.068439	C11H23NO2S4	H+	-0.453	0.0338	9
324.2043745	324.20467	C13H30N3O4P	H+	-0.911	0.0340	2
307.1751152	307.175131	C14H26O7	H+	-0.051	0.0352	2
481.3019638	481.302062	C24H40N4O6	H+	-0.204	0.0363	7
324.2016724	324.20168	C14H29NO7	H+	-0.023	0.0367	1
480.2624967	480.262566	C22H41NO8S	H+	-0.144	0.0380	5
388.0775975	388.077543	C17H13N3O8	H+	0.140	0.0392	13
478.2644759	478.264675	C22H39NO10	H+	-0.416	0.0398	4
356.047405	356.047704	C12H21NO3S4	H+	-0.840	0.0400	11
424.2540555	424.25411	C19H37NO9	H+	-0.129	0.0403	2
316.211798	316.21185	C16H29NO5	H+	-0.164	0.0406	3
338.2537388	338.253715	C16H35NO6	H+	0.070	0.0406	0
374.290015	374.2901	C20H39NO5	H+	-0.227	0.0407	2

308.1703467	308.17038	C13H25NO7	H+	-0.108	0.0410	2
243.15908	243.159086	C13H22O4	H+	-0.025	0.0421	3
386.2536904	386.253715	C20H35NO6	H+	-0.064	0.0422	4
320.1703762	320.17038	C14H25NO7	H+	-0.012	0.0424	3
478.2437525	478.243545	C25H35NO8	H+	0.434	0.0427	9
194.1387075	194.138685	C8H19NO4	H+	0.116	0.0430	0
348.1656295	348.165295	C15H25NO8	H+	0.961	0.0435	4
648.3800092	648.380586	C31H58N3O7SP	H+	-0.890	0.0436	8
384.222918	384.22281	C16H33NO9	H+	0.281	0.0436	1
348.1593393	348.15942	C22H21NO3	H+	-0.232	0.0437	13
684.1980557	684.198175	C26H37NO20	H+	-0.174	0.0439	9
456.1543583	456.154747	C18H25N5O7S	H+	-0.852	0.0450	11
319.211525	319.211516	C16H30O6	H+	0.028	0.0451	2
666.3905069	666.391151	C31H60N3O8SP	H+	-0.967	0.0452	7
344.243275	344.24315	C18H33NO5	H+	0.363	0.0455	3
522.2910491	522.29089	C24H43NO11	H+	0.305	0.0456	4
393.2352142	393.235244	C17H36N4O2S2	H+	-0.076	0.0458	6
428.2490628	428.249025	C18H37NO10	H+	0.088	0.0458	1
280.1754183	280.175465	C12H25NO6	H+	-0.167	0.0459	1
338.2564044	338.256705	C15H36N3O3P	H+	-0.889	0.0462	1
422.29412	422.294338	C22H47NS3	H+	-0.516	0.0462	6
348.2378264	348.238065	C17H33NO6	H+	-0.685	0.0463	2
500.26464	500.26428	C28H37NO7	H+	0.720	0.0465	11
359.0567954	359.056886	C12H23O4S3P	H+	-0.252	0.0465	9
321.2391138	321.239266	C16H36N2S2	H+	-0.474	0.0469	4
371.0834154	371.083755	C14H26O3S4	H+	-0.915	0.0480	10
318.1911432	318.191115	C15H27NO6	H+	0.089	0.0482	3
556.2960686	556.29637	C24H45NO13	H+	-0.542	0.0483	3
348.2315091	348.231539	C16H33N3O3S	H+	-0.086	0.0487	4
404.227804	404.227895	C19H33NO8	H+	-0.225	0.0496	4
348.1436285	348.143514	C10H25N3O8S	H+	0.329	0.0496	2
457.152692	457.152986	C16H25N8O4SP	H+	-0.643	0.0499	12
440.2491375	440.249025	C19H37NO10	H+	0.256	0.0499	2

Table 7: Metabolic REMD list.

Reaction	ΔMass	H	C	O	N	S	P	Na+
amino-function exchanged by hydroxyl function	0.984016	-1	0	1	-1	0	0	0
deamination	1.031634	3	0	-1	1	0	0	0
hydrolysis of acetic acid and consecutive carboxylation| decarboxylative condensation	1.979265	-2	-1	1	0	0	0	0
(de-)hydrogenation	2.01565	2	0	0	0	0	0	0
glyoxylic acid| decarboxylative condensation	12	0	1	0	0	0	0	0
glycine| decarboxylative condensation	13.031634	3	1	-1	1	0	0	0
methylation	14.01565	2	1	0	0	0	0	0
H2N● - H +neutral/reductive deamination	15.010899	1	0	0	1	0	0	0
methylation on tertiary N (like in N-trimethyl-lysine)	15.023475	3	1	0	0	0	0	0
exchange of O with S	15.977156	0	0	-1	0	1	0	0
(de-)hydroxylation	15.994915	0	0	1	0	0	0	0
hydrolysis/condensation	18.010565	2	0	1	0	0	0	0
Self	0	0	0	0	0	0	0	1
C=O insertion like in biotin synthesis or hydroxymethyl-transfer	25.979265	-2	1	1	0	0	0	0
pyruvic acid| decarboxylative condensation	26.01565	2	2	0	0	0	0	0
butanoic acid| decarboxylative condensation	26.052035	6	3	-1	0	0	0	0
formimino transfer	27.010899	1	1	0	1	0	0	0

alanineǀ decarboxylative condensation	27.047284	5	2	-1	1	0	0	0
formyl transfer	27.994915	0	1	1	0	0	0	0
C2H4	28.0313	4	2	0	0	0	0	0
NO● - H +neutral (nitrosylation)	28.990164	-1	0	1	1	0	0	0
thio-heteroatom	29.956421	-2	0	0	0	1	0	0
hydroxymethyl transfer	30.010565	2	1	1	0	0	0	0
thiolation	31.972071	0	0	0	0	1	0	0
hydro-peroxidation	31.98983	0	0	2	0	0	0	0
hydroxypyruvic acidǀ decarboxylative condensation	42.010565	2	2	1	0	0	0	0
guanidyl group transfer	42.021798	2	1	0	2	0	0	0
carbamoyl or isocyainde transferǀ	43.005814	1	1	1	1	0	0	0
serineǀ decarboxylative condensation	43.042199	5	2	0	1	0	0	0
(de-)carboxylation	43.98983	0	1	2	0	0	0	0
pyruvic acidǀ decarboxylative addition	44.026215	4	2	1	0	0	0	0
nitration (+NO2 -H)	44.985079	-1	0	2	1	0	0	0
phospholytic decarboxylation	35.976502	1	-1	1	0	0	1	0
prolineǀ decarboxylative condensation	53.062934	7	4	-1	1	0	0	0
α-ketoisovaleric acidǀ decarboxylative condensation	54.04695	6	4	0	0	0	0	0
hexanoic acidǀ decarboxylative condensation	54.083335	10	5	-1	0	0	0	0
valineǀ decarboxylative condensation	55.078584	9	4	-1	1	0	0	0
glyoxylic acidǀ condensation	55.98983	0	2	2	0	0	0	0

3-hydroxy-2-oxobutanoic acidl decarboxylative condensation	56.026215	4	3	1	0	0	0	0
glycinel condensations	57.021464	3	2	1	1	0	0	0
threoninel decarboxylative condensation	57.057849	7	3	0	1	0	0	0
3-mercaptopyruvatel decarboxylative condensation	57.987721	2	2	0	0	1	0	0
cysteinel decarboxylative condensation	59.019355	5	2	-1	1	1	0	0
glycinel condensations on hydrogenated carbonyls	59.037114	5	2	1	1	0	0	0
hydroxypyruvic acidl decarboxylative addition	60.02113	4	2	2	0	0	0	0
prenylation	68.0626	8	5	0	0	0	0	0
4-amino-2,4-dioxobutanoic acidl decarboxylative condensation	69.021464	3	3	1	1	0	0	0
5-amino-2-oxopentanoic acidl decarboxylative condensation	69.057849	7	4	0	1	0	0	0
leucine/isoleucinel decarboxylative condensation	69.094234	11	5	-1	1	0	0	0
2-ketosuccinatel decarboxylative condensation	70.00548	2	3	2	0	0	0	0
butanoic acidl l condensation	70.041865	6	4	1	0	0	0	0
asparaginel decarboxylative condensation	70.053098	6	3	0	2	0	0	0

ornithine| decarboxylative condensation	70.089483	10	4	-1	2	0	0	0
alanine| condensations	71.037114	5	3	1	1	0	0	0
oxalate| condensation	71.984745	0	2	3	0	0	0	0
α-ketoisovaleric acid| decarboxylative addition	72.057515	8	4	1	0	0	0	0
EC 4.1.99.1 Tryptophanase	73.016379	3	2	2	1	0	0	0
alanine| condensations on hydrogenated carbonyls	73.052764	7	3	1	1	0	0	0
3-hydroxy-2-oxobutanoic acid| decarboxylative addition	74.03678	6	3	2	0	0	0	0
3-mercaptopyruvate| decarboxylative addition	75.998286	4	2	1	0	1	0	0
direct sulfonation and sulfonic anhydride (-H2O)	79.956816	0	0	3	0	1	0	0
(de-)phosphorylation	79.966332	1	0	3	0	0	1	0
octanoic acid| decarboxylative condensation	82.114635	14	7	-1	0	0	0	0
2-keto-glutaramic acid| decarboxylative condensation	83.037114	5	4	1	1	0	0	0
2-keto-6-aminocaproate| decarboxylative condensation	83.073499	9	5	0	1	0	0	0
2-ketoglutarate| decarboxylative condensation	84.02113	4	4	2	0	0	0	0
adipate| decarboxylative condensation	84.057515	8	5	1	0	0	0	0

glutamine| decarboxylative condensation	84.068748	8	4	0	2	0	0	0
lysine| decarboxylative condensation	84.105133	12	5	-1	2	0	0	0
glutamic acid| decarboxylative condensation	85.052764	7	4	1	1	0	0	0
hydroxypyruvic acid| condensation	86.000395	2	3	3	0	0	0	0
2-oxo-4-methylthiobutanoic acid| decarboxylative condensation	86.019021	6	4	0	0	1	0	0
2-ketohexanoic acid| decarboxylative addition	86.073165	10	5	1	0	0	0	0
serine| condensations	87.032029	5	3	2	1	0	0	0
methionine| decarboxylative condensation	87.050655	9	4	-1	1	1	0	0
5-amino-2-oxopentanoic acid| decarboxylative addition	87.068414	9	4	1	1	0	0	0
2-ketosuccinate| decarboxylative addition	88.016045	4	3	3	0	0	0	0
serine| condensations on hydrogenated carbonyls	89.047679	7	3	2	1	0	0	0
2-oxo-3-(1H-imidazol-4-yl)propansäure| decarboxylative condensation	92.037448	4	5	0	2	0	0	0
histidine| decarboxylative condensation	93.069082	7	5	-1	3	0	0	0
proline| condensations	97.052764	7	5	1	1	0	0	0
α-ketoisovaleric acid|	98.03678	6	5	2	0	0	0	0

condensation								
hexanoic acid| condensation	98.073165	10	6	1	0	0	0	0
valine or proline - condensation on hydrogenated carbonyls| condensations	99.068414	9	5	1	1	0	0	0
3-hydroxy-2-oxobutanoic acid| condensation	100.01605	4	4	3	0	0	0	0
threonine| condensations	101.04768	7	4	2	1	0	0	0
valine| condensations on hydrogenated carbonyls	101.08406	11	5	1	1	0	0	0
3-mercaptopyruvate| condensation	101.97755	2	3	2	0	1	0	0
2-ketoglutarate| decarboxylative addition	102.0317	6	4	3	0	0	0	0
phenylpyruvic acid| decarboxylative condensation	102.04695	6	8	0	0	0	0	0
cysteine| condensations	103.00919	5	3	1	1	1	0	0
threonine| condensations on hydrogenated carbonyls	103.06333	9	4	2	1	0	0	0
phenylalanine| decarboxylative condensation	103.07858	9	8	-1	1	0	0	0
2-oxo-4-methylthiobutanoic acid| decarboxylative addition	104.02959	8	4	1	0	1	0	0
cysteine| condensations on hydrogenated carbonyls	105.02484	7	3	1	1	1	0	0
taurine| condensations	107.0041	5	2	2	1	1	0	0

taurine| condensations on hydrogenated carbonyls	109.01975	7	2	2	1	1	0	0
2-oxo-3-(1H-imidazol-4-yl)propansäure| decarboxylative addition	110.04801	6	5	1	2	0	0	0
decanoic acid| decarboxylative condensation	110.14594	18	9	-1	0	0	0	0
2-oxoarginine| decarboxylative condensation	111.07965	9	5	0	3	0	0	0
2-ketohexanoic acid| condensation	112.05243	8	6	2	0	0	0	0
suberate| decarboxylative condensation	112.08882	12	7	1	0	0	0	0
arginine| decarboxylative condensation	112.11128	12	5	-1	4	0	0	0
4-amino-2,4-dioxobutanoic acid| condensation	113.01129	3	4	3	1	0	0	0
5-amino-2-oxopentanoic acid| condensation	113.04768	7	5	2	1	0	0	0
leucine/Isoleucine| condensations	113.08406	11	6	1	1	0	0	0
2-ketosuccinate| condensation	113.99531	2	4	4	0	0	0	0
glutarate| condensation	114.0317	6	5	3	0	0	0	0
asparagine| condensations	114.04293	6	4	2	2	0	0	0
ornithine| condensations	114.07931	10	5	1	2	0	0	0
aspartic acid| condensations	115.02694	5	4	3	1	0	0	0

leucine/isoleucine| condensations on hydrogenated carbonyls	115.09971	13	6	1	1	0	0	0
asparagine| condensations on hydrogenated carbonyls	116.05858	8	4	2	2	0	0	0
ornithine| condensations on hydrogenated carbonyls	116.09496	12	5	1	2	0	0	0
aspartic acid| condensations on hydrogenated carbonyls	117.04259	7	4	3	1	0	0	0
4-hydroxyphenylpyruvic acid| decarboxylative condensation	118.04187	6	8	1	0	0	0	0
tyrosine| decarboxylative condensation	119.0735	9	8	0	1	0	0	0
phenylpyruvic acid| decarboxylative addition	120.05752	8	8	1	0	0	0	0
phosphoethanolamine	123.00853	6	2	3	1	0	1	0
octanoic acid| condensation	126.10447	14	8	1	0	0	0	0
2-keto-glutaramic acid| condensation	127.02694	5	5	3	1	0	0	0
2-keto-6-aminocaproate| condensation	127.06333	9	6	2	1	0	0	0
2-ketoglutarate| condensation	128.01096	4	5	4	0	0	0	0
adipate| condensation	128.04735	8	6	3	0	0	0	0
glutamine| condensations	128.05858	8	5	2	2	0	0	0
lysine| condensations	128.09496	12	6	1	2	0	0	0

glutamic acid| condensations	129.04259	7	5	3	1	0	0	0
2-oxoarginine| decarboxylative addition	129.09021	11	5	1	3	0	0	0
2-oxo-4-methylthiobutanoic acid| condensation	130.00885	6	5	2	0	1	0	0
glutamine| condensations on hydrogenated carbonyls	130.07423	10	5	2	2	0	0	0
lysine| condensations on hydrogenated carbonyls	130.11061	14	6	1	2	0	0	0
methionine| condensations	131.04049	9	5	1	1	1	0	0
glutamic acid| condensations on hydrogenated carbonyls	131.05824	9	5	3	1	0	0	0
methionine| condensations on hydrogenated carbonyls	133.05614	11	5	1	1	1	0	0
2-oxo-3-(1H-imidazol-4-yl)propansäure| condensation	136.02728	4	6	2	2	0	0	0
4-hydroxyphenylpyruvic acid| decarboxylative addition	136.05243	8	8	2	0	0	0	0
di-prenylation	136.1252	16	10	0	0	0	0	0
histidine| condensations	137.05891	7	6	1	3	0	0	0
dodecanoic acid| decarboxylative condensation	138.17724	22	11	-1	0	0	0	0
histidine| condensations on hydrogenated carbonyls	139.07456	9	6	1	3	0	0	0

sebacic acid| decarboxylative condensation	140.12012	16	9	1	0	0	0	0
indole pyruvic acid| decarboxylative condensation	141.05785	7	10	0	1	0	0	0
pimelate| condensation	142.063	10	7	3	0	0	0	0
tryptophan| decarboxylative condensation	142.08948	10	10	-1	2	0	0	0
phenylpyruvic acid| condensation	146.03678	6	9	2	0	0	0	0
phenylalanine| condensations	147.06841	9	9	1	1	0	0	0
phenylalanine| condensations on hydrogenated carbonyls	149.08406	11	9	1	1	0	0	0
glycerol-3-phosphate	154.00311	7	3	5	0	0	1	0
decanoic acid| condensation	154.13577	18	10	1	0	0	0	0
2-oxoarginine| condensation	155.06948	9	6	2	3	0	0	0
suberate| condensation	156.07865	12	8	3	0	0	0	0
arginine| condensations	156.10111	12	6	1	4	0	0	0
arginine| condensations on hydrogenated carbonyls	158.11676	14	6	1	4	0	0	0
indole pyruvic acid| decarboxylative addition	159.06841	9	10	1	1	0	0	0
4-hydroxyphenylpyruvic acid| condensation	162.0317	6	9	3	0	0	0	0
glucose	162.05283	10	6	5	0	0	0	0
tyrosine| condensations	163.06333	9	9	2	1	0	0	0

tyrosine| condensations on hydrogenated carbonyls	165.07898	11	9	2	1	0	0	0
tetradecanoic acid| decarboxylative condensation	166.20854	26	13	-1	0	0	0	0
phosphatidylserine	166.99836	6	3	5	1	0	1	0
azelaic acid| condensation	170.0943	14	9	3	0	0	0	0
glucuronidation	176.03209	8	6	6	0	0	0	0
dodecanoic acid| condensation	182.16707	22	12	1	0	0	0	0
sebacic acid| condensation	184.10995	16	10	3	0	0	0	0
indole pyruvic acid| condensation	185.04768	7	11	2	1	0	0	0
tryptophan| condensations	186.07931	10	11	1	2	0	0	0
tryptophan| condensations on hydrogenated carbonyls	188.09496	12	11	1	2	0	0	0
hexadecanoic acid| decarboxylative condensation	194.23984	30	15	-1	0	0	0	0
tri-prenylation	204.1878	24	15	0	0	0	0	0
tetradecanoic acid| condensation	210.19837	26	14	1	0	0	0	0
ribose-5-phosphate	212.00859	9	5	7	0	0	1	0
hexadecanoic acid| condensation	238.22967	30	16	1	0	0	0	0
phosphatidylinositol	242.01916	11	6	8	0	0	1	0
tetra-prenylation	272.2504	32	20	0	0	0	0	0
phosphatidylcholine head group	239.09226	18	8	5	1	0	1	0
phosphorylcholine	166.06331	13	5	3	1	0	1	0
self	1E-09	0	0	0	0	0	0	0

Table 8: Markers of HuMet Block E4|E6.

E-Block	Exact Mass	Formula	Cyclomatic Number	MassTRIX
E4|E6	155.04271	C4H8N2O3	2	L-asparagine
E4|E6	144.08078	C10H9N	7	2-naphthylamine
E4|E6	164.97825	C4H7OP3	6	0
E4|E6	169.12231	C10H16O2	3	6-oxocineole
E4|E6	171.06519	C8H10O4	4	3,4-dihydroxyphenylethyleneglycol ([M+H]+)
E4|E6	174.18524	C10H23NO	0	0
E4|E6	175.08996	C7H14N2OS	4	0
E4|E6	197.07843	C8H14O4	2	suberic acid
E4|E6	178.12264	C11H15NO	5	phenmetrazine ([M+H]+)
E4|E6	180.06552	C9H9NO3	6	hippurate
E4|E6	180.13829	C11H17NO	4	mexiletine ([M+H]+)
E4|E6	205.06826	C6H14O6	0	mannitol
E4|E6	185.11722	C10H16O3	3	5-exo-hydroxy-1,2-campholide ([M+H]+)
E4|E6	187.12635	C9H18N2S	4	0
E4|E6	189.15975	C9H20N2O2	1	7,8-diaminononanoate ([M+H]+)
E4|E6	193.15869	C13H20O	4	alpha-ionone ([M+H]+)
E4|E6	195.13796	C12H18O2	4	4-hexyloxyphenol ([M+H]+)
E4|E6	197.16484	C11H20N2O	3	0
E4|E6	198.14886	C11H19NO2	3	0
E4|E6	199.13287	C11H18O3	3	0
E4|E6	199.1441	C10H18N2O2	3	0
E4|E6	200.16451	C11H21NO2	2	0
E4|E6	201.12337	C9H16N2O3	3	0
E4|E6	201.12739	C14H16O	7	0
E4|E6	201.14852	C11H20O3	2	2-hydroxy-10-undecenoic acid [hydroxy fatty acids [FA0105]] ([M+H]+)
E4|E6	223.16685	C12H24O2	1	dodecanoic acid
E4|E6	208.16959	C13H21NO	4	luciduline ([M+H]+)
E4|E6	209.11722	C12H16O3	5	benzyl (2R,3S)-2-methyl-3-

				hydroxybutanoate ([M+H]+)
E4|E6	209.15361	$C_{13}H_{20}O_2$	4	4-heptyloxyphenol
E4|E6	211.13287	$C_{12}H_{18}O_3$	4	(-)-jasmonic acid
E4|E6	211.15534	$C_{10}H_{18}N_4O$	4	0
E4|E6	212.12812	$C_{11}H_{17}NO_3$	4	mescaline ([M+H]+)
E4|E6	212.16451	$C_{12}H_{21}NO_2$	3	elaeokanine C ([M+H]+)
E4|E6	214.12264	$C_{14}H_{15}NO$	8	0
E4|E6	237.15735	$C_{11}H_{22}N_2O_2$	2	0
E4|E6	216.19581	$C_{12}H_{25}NO_2$	1	12-amino-dodecanoic acid [amino fatty acids [FA0110]] ([M+H]+)
E4|E6	217.10705	$C_{10}H_{16}O_5$	3	0
E4|E6	239.12538	$C_{11}H_{20}O_4$	2	undecanedioic acid
E4|E6	218.13869	$C_{10}H_{19}NO_4$	2	O-propanoylcarnitine
E4|E6	219.17434	$C_{15}H_{22}O$	5	alpha-sinensal ([M+H]+)
E4|E6	221.11722	$C_{13}H_{16}O_3$	6	precocene 2
E4|E6	223.09649	$C_{12}H_{14}O_4$	6	apiole ([M+H]+)
E4|E6	223.16926	$C_{14}H_{22}O_2$	4	rishitin ([M+H]+)
E4|E6	225.18491	$C_{14}H_{24}O_2$	3	5,8-tetradecadienoic acid
E4|E6	226.14377	$C_{12}H_{19}NO_3$	4	terbutaline ([M+H]+)
E4|E6	229.14344	$C_{12}H_{20}O_4$	3	traumatic acid
E4|E6	230.13869	$C_{11}H_{19}NO_4$	3	butenylcarnitine [unclassified substance] ([M+H]+)
E4|E6	230.17507	$C_{12}H_{23}NO_3$	2	N-decanoylglycine [carboxylic acid] ([M+H]+)
E4|E6	231.1227	$C_{11}H_{18}O_5$	3	0
E4|E6	233.11722	$C_{14}H_{16}O_3$	7	kavapyrone ([M+H]+)
E4|E6	233.14958	$C_{10}H_{20}N_2O_4$	2	spermic acid 2
E4|E6	239.12779	$C_{13}H_{18}O_4$	5	0
E4|E6	239.16417	$C_{14}H_{22}O_3$	4	7-oxo-11E,13-tetradecadienoic acid [oxo fatty acids [FA0106]] ([M+H]+)
E4|E6	239.20056	$C_{15}H_{26}O_2$	3	centarol ([M+H]+)
E4|E6	240.15942	$C_{13}H_{21}NO_3$	4	isoetharine ([M+H]+)
E4|E6	240.19581	$C_{14}H_{25}NO_2$	3	0
E4|E6	242.17507	$C_{13}H_{23}NO_3$	3	valeroidine ([M+H]+)

E4|E6	244.15434	$C_{12}H_{21}NO_4$	3	tiglylcarnitine [cation] ([M+H]+)
E4|E6	246.16999	$C_{12}H_{23}NO_4$	2	2-methylbutyroylcarnitine
E4|E6	271.07883	$C_{10}H_{16}O_7$	3	0
E4|E6	249.14852	$C_{15}H_{20}O_3$	6	1,2-dihydrosantonin ([M+H]+)
E4|E6	255.23186	$C_{16}H_{30}O_2$	2	(9Z)-hexadecenoic acid
E4|E6	256.19072	$C_{14}H_{25}NO_3$	3	0
E4|E6	256.26349	$C_{16}H_{33}NO$	1	palmitic amide
E4|E6	257.16484	$C_{16}H_{20}N_2O$	8	chanoclavine-I ([M+H]+)
E4|E6	259.19039	$C_{14}H_{26}O_4$	2	2,3,4-trioxycyclopentanone ([M+H]+)
E4|E6	262.23767	$C_{14}H_{31}NO_3$	0	0
E4|E6	265.11828	$C_{13}H_{16}N_2O_4$	7	alpha-N-phenylacetyl-L-glutamine ([M+H]+)
E4|E6	265.17982	$C_{16}H_{24}O_3$	5	dehydrojuvabione ([M+H]+)
E4|E6	267.23186	$C_{17}H_{30}O_2$	3	7-heptadecynoic acid [Unsaturated fatty acids [FA0103]] ([M+H]+)
E4|E6	268.19072	$C_{15}H_{25}NO_3$	4	metoprolol ([M+H]+)
E4|E6	269.17474	$C_{15}H_{24}O_4$	4	0
E4|E6	269.18597	$C_{14}H_{24}N_2O_3$	4	0
E4|E6	269.21112	$C_{16}H_{28}O_3$	3	(1R,2R)-3-oxo-2-pentyl-cyclopentanehexanoic acid [12-oxophytodienoic acid metabolites [FA0201]] ([M+H]+)
E4|E6	270.15473	$C_{10}H_{23}NO_7$	0	0
E4|E6	270.20637	$C_{15}H_{27}NO_3$	3	0
E4|E6	271.20162	$C_{14}H_{26}N_2O_3$	3	0
E4|E6	271.26316	$C_{17}H_{34}O_2$	1	methyl palmitate ([M+H]+)
E4|E6	272.20089	$C_{18}H_{25}NO$	7	dextromethorphan ([M+H]+)
E4|E6	272.23325	$C_{14}H_{29}N_3O_2$	2	0
E4|E6	272.25841	$C_{16}H_{33}NO_2$	1	2R-aminohexadecanoic acid [amino fatty acids [FA0110]] ([M+H]+)
E4|E6	277.12818	$C_{12}H_{20}O_7$	3	0
E4|E6	277.21621	$C_{18}H_{28}O_2$	5	R replaced by H in steryl ester
E4|E6	278.15982	$C_{12}H_{23}NO_6$	2	0

E4|E6	279.19547	$C_{17}H_{26}O_3$	5	[6]-paradol
E4|E6	279.23186	$C_{18}H_{30}O_2$	4	(6Z,9Z,12Z)-octadecatrienoic acid
E4|E6	280.15434	$C_{15}H_{21}NO_4$	6	metalaxyl ([M+H]+)
E4|E6	280.26349	$C_{18}H_{33}NO$	3	linoleamide
E4|E6	304.18832	$C_{16}H_{27}NO_3$	4	0
E4|E6	283.21285	$C_{14}H_{26}N_4O_2$	4	0
E4|E6	283.22677	$C_{17}H_{30}O_3$	3	6-oxabicyclo[3.1.0]hexane-2-undecanoic acid methyl ester ([M+H]+)
E4|E6	284.17038	$C_{11}H_{25}NO_7$	0	0
E4|E6	284.22202	$C_{16}H_{29}NO_3$	3	0
E4|E6	285.16965	$C_{15}H_{24}O_5$	4	0
E4|E6	285.24242	$C_{17}H_{32}O_3$	2	2-methoxy-5Z-hexadecenoic acid
E4|E6	286.1649	$C_{14}H_{23}NO_5$	4	0
E4|E6	286.20129	$C_{15}H_{27}NO_4$	3	2-octenoylcarnitine [cation] ([M+H]+)
E4|E6	287.14892	$C_{14}H_{22}O_6$	4	0
E4|E6	288.21694	$C_{15}H_{29}NO_4$	2	L-octanoylcarnitine ([M+H]+)
E4|E6	293.17474	$C_{17}H_{24}O_4$	6	trichodermin ([M+H]+)
E4|E6	293.19587	$C_{14}H_{28}O_6$	1	0
E4|E6	293.21112	$C_{18}H_{28}O_3$	5	12-OPDA
E4|E6	294.20637	$C_{17}H_{27}NO_3$	5	(+/-)-5-[(tert-butylamino)-2'-hydroxypropoxy]-1,2,3,4-tetrahydro-1-naphthol
E4|E6	296.22202	$C_{17}H_{29}NO_3$	4	0
E4|E6	319.11521	$C_{15}H_{20}O_6$	6	vomitoxin
E4|E6	297.27881	$C_{19}H_{36}O_2$	2	oleic acid methyl ester
E4|E6	298.27406	$C_{18}H_{35}NO_2$	2	3-ketosphingosine ([M+H]+)
E4|E6	298.34683	$C_{20}H_{43}N$	0	0
E4|E6	299.06176	$C_{10}H_{18}O_6S_2$	6	0
E4|E6	301.28495	$C_{17}H_{36}N_2O_2$	1	0
E4|E6	302.1962	$C_{15}H_{27}NO_5$	3	0
E4|E6	324.94936	$C_7H_{12}O_5S_2P_2$	9	0
E4|E6	303.23186	$C_{20}H_{30}O_2$	6	abietate

E4|E6	304.17547	C14H25NO6	3	pimelylcarnitine [cation] ([M+H]+)
E4|E6	307.09986	C16H18O4S	10	0
E4|E6	307.22677	C19H30O3	5	oxandrolone ([M+H]+)
E4|E6	310.31044	C20H39NO	2	0
E4|E6	311.22169	C18H30O4	4	13(S)-HPOT
E4|E6	311.23292	C17H30N2O3	4	0
E4|E6	311.25807	C19H34O3	3	methoprene
E4|E6	312.21694	C17H29NO4	4	2-trans,4-cis-decadienoylcarnitine [cation] ([M+H]+)
E4|E6	312.32609	C20H41NO	1	0
E4|E6	314.1962	C16H27NO5	4	heliotrine
E4|E6	314.24382	C16H31N3O3	3	0
E4|E6	318.24276	C20H31NO2	6	17beta-hydroxy-4,17-dimethyl-4-azaandrost-5-en-3-one ([M+H]+)
E4|E6	321.24242	C20H32O3	5	(15S)-15-hydroxy-5,8,11-cis-13-trans-eicosatetraenoate
E4|E6	321.31519	C22H40O	3	0
E4|E6	323.25807	C20H34O3	4	2alpha-(hydroxymethyl)-5alpha-androstane-3beta,17beta-diol ([M+H]+)
E4|E6	325.11293	C12H20O10	3	bis-D-fructose 2',1:2,1'-dianhydride
E4|E6	325.27372	C20H36O3	3	alchornoic acid
E4|E6	328.1966	C13H29NO8	0	0
E4|E6	328.24824	C18H33NO4	3	10-nitro-9E-octadecenoic acid [nitro fatty acids [FA0112]] ([M+H]+)
E4|E6	331.28429	C19H38O4	1	MG(0:0/16:0/0:0)
E4|E6	334.31044	C22H39NO	4	0
E4|E6	335.19819	C16H31O5P	3	0
E4|E6	335.22169	C20H30O4	6	12-keto-leukotriene B4
E4|E6	339.28937	C21H38O3	3	0
E4|E6	339.32576	C22H42O2	2	(13Z)-docosenoic acid
E4|E6	343.20883	C14H26N6O4	5	0
E4|E6	348.23807	C17H33NO6	2	0
E4|E6	349.31011	C23H40O2	4	20:3(5Z,9Z,17Z)(11Me,15Me,19Me)

E4|E6	356.35231	C22H45NO2	1	eicosanoyl-EA
E4|E6	358.2422	C12H27N11O2	5	0
E4|E6	358.27003	C18H35N3O4	3	leucyl-leucyl-norleucine ([M+H]+)
E4|E6	363.21997	C18H34O5S	4	0
E4|E6	363.25195	C16H35N4O3P	3	0
E4|E6	363.3105	C20H42O5	0	0
E4|E6	388.18728	C17H28N5O2P	8	0
E4|E6	366.37304	C24H47NO	2	0
E4|E6	367.21152	C20H30O6	6	20-COOH-leukotriene B4
E4|E6	369.35158	C27H44	6	3-deoxyvitamin D3
E4|E6	370.29519	C21H39NO4	3	cis-5-tetradecenoylcarnitine [cation] ([M+H]+)
E4|E6	371.32682	C21H42N2O3	2	0
E4|E6	372.34722	C22H45NO3	1	0
E4|E6	376.34214	C21H45NO4	0	0
E4|E6	382.2588	C21H35NO5	5	0
E4|E6	384.27445	C21H37NO5	4	3-hydroxy-5, 8-tetradecadiencarnitine [cation] ([M+H]+)
E4|E6	387.08732	C13H23O9SP	6	0
E4|E6	387.15528	C13H30N4O3S3	7	0
E4|E6	387.18359	C19H30O6S	7	0
E4|E6	390.21225	C18H31NO8	4	0
E4|E6	392.17626	C13H29NO12	0	0
E4|E6	398.36287	C24H47NO3	2	behenoylglycine [carboxylic acid] ([M+H]+)
E4|E6	400.37852	C24H49NO3	1	0
E4|E6	400.41491	C25H53NO2	0	0
E4|E6	406.13825	C16H28N3O3S2P	10	0
E4|E6	409.40401	C27H52O2	2	(+)-C27-phthienoic acid
E4|E6	413.32615	C24H44O5	3	0
E4|E6	420.35577	C21H41N9	6	0

E4|E6	423.3945	C26H50N2O2	3	0
E4|E6	433.33124	C27H44O4	6	gitogenin ([M+H]+)
E4|E6	433.40401	C29H52O2	4	29:3(5Z,9Z,23Z)
E4|E6	437.43531	C29H56O2	2	mycolipenic acid (C29)
E4|E6	441.39384	C27H52O4	2	MG(0:0/24:1(15Z)/0:0)
E4|E6	475.31156	C21H40N8O3	6	0
E4|E6	453.34353	C24H44N4O4	5	0
E4|E6	453.366	C22H44N8O2	5	0
E4|E6	458.32161	C25H48NO2SP	6	0
E4|E6	459.35141	C19H42N10O3	4	0
E4|E6	460.1966	C24H29NO8	11	0
E4|E6	464.37344	C28H49NO4	5	0
E4|E6	466.42547	C29H55NO3	3	0
E4|E6	471.3541	C24H46N4O5	4	0
E4|E6	484.39965	C28H53NO5	3	0
E4|E6	520.33302	C22H43N9O4	6	0
E4|E6	500.28226	C21H45N3O6S2	5	0
E4|E6	566.4276	C30H55N5O5	6	0

Table 9: Markers HuMet Block E5.

E-Block	Exact Mass	Formula	Cyclomatic Number	MassTRIX
E5	181.07699	C7H14N2S	4	0
E5	163.07536	C10H10O2	6	cis-1,2-dihydronaphthalene-1,2-diol
E5	163.11174	C11H14O	5	4,10-undecadiynal [fatty aldehydes [FA06]] ([M+H]+)
E5	198.18524	C12H23NO	2	0
E5	199.16926	C12H22O2	2	(-)-menthyl acetate
E5	217.08489	C7H13N4O2P	5	0
E5	217.17982	C12H24O3	1	12-hydroxydodecanoic acid

E5	221.08421	C9H16O4S	4	0
E5	223.13287	C13H18O3	5	dehydrovomifoliol
E5	258.1312	C10H21NO5	1	0
E5	238.1649	C10H23NO5	0	0
E5	242.13869	C12H19NO4	4	N-(3-oxooctanoyl)homoserine lactone
E5	249.16141	C12H25O3P	2	0
E5	251.16417	C15H22O3	5	arbusculin A ([M+H]+)
E5	251.1853	C12H26O5	0	0
E5	253.25259	C17H32O	2	2,6,8,12-tetramethyl-2,4-tridecadien-1-ol [fatty alcohols [FA05]] ([M+H]+)
E5	277.18865	C14H26N2O2	3	0
E5	257.24751	C16H32O2	1	hexadecanoic acid
E5	267.1227	C14H18O5	6	0
E5	273.16965	C14H24O5	3	0
E5	274.27406	C16H35NO2	0	hexadecasphinganine ([M+H]+)
E5	287.22169	C16H30O4	2	2,3-dihydroxycyclopentane-undecanoic acid ([M+H]+)
E5	288.28971	C17H37NO2	0	C17 sphinganine
E5	321.24002	C18H34O3	2	2-oxooctadecanoic acid
E5	304.09445	C12H18NO6P	6	0
E5	309.21892	C15H33O4P	1	0
E5	318.30027	C18H39NO3	0	phytosphingosine
E5	329.26864	C19H36O4	2	MG(0:0/16:1(9Z)/0:0)
E5	330.99872	C9H16O5S2P2	9	0
E5	332.24315	C17H33NO5	2	0
E5	333.29994	C19H40O4	0	0
E5	340.28462	C20H37NO3	3	oleoyl glycine
E5	356.27954	C20H37NO4	3	0
E5	357.16965	C21H24O5	10	rutamarin ([M+H]+)
E5	372.21694	C22H29NO4	9	0

E5	372.25669	C20H37NO3S	5	0
E5	413.26623	C24H38O4	6	bis(2-ethylhexyl)phthalate
E5	397.33124	C24H44O4	3	0
E5	398.23259	C24H31NO4	10	0
E5	405.22448	C19H28N6O4	9	0
E5	414.35779	C24H47NO4	2	heptadecanoyl carnitine
E5	425.41015	C26H52N2O2	2	0
E5	463.14607	C17H33N2O5S2P	8	0
E5	451.4258	C28H54N2O2	3	0
E5	454.46186	C29H59NO2	1	0
E5	611.18864	C23H40N4O7S2P2	13	0

Table 10: Markers HuMet Block E8

E-Block	Exact Mass	Formula	Cyclomatic Number	MassTRIX
E8	128.03182	C3H7NO3	1	L-serine
E8	151.08418	C6H12N2O	2	L-lysine 1,6-lactam
E8	136.06177	C5H5N5	6	adenine
E8	165.05222	C7H10O3	3	4-oxocyclohexanecarboxylate ([M+H]+)
E8	174.99531	C4H9O2SP	4	0
E8	157.12231	C9H16O2	2	nonane-4,6-dione ([M+H]+)
E8	183.06278	C7H12O4	2	6-carboxyhexanoate
E8	183.09917	C8H16O3	1	ethyl (R)-3-hydroxyhexanoate ([M+H]+)
E8	170.0924	C7H11N3O2	4	N(pi)-methyl-L-histidine
E8	172.09682	C8H13NO3	3	crotanecine ([M+H]+)
E8	201.0886	C11H14O2	5	eugenol methyl ether
E8	203.05261	C6H12O6	1	D-glucose
E8	207.13555	C11H20O2	2	gamma-undecalactone
E8	202.155	C9H19N3O2	2	0

E8	227.17702	C15H24	4	pentalenene ([M+H]+)
E8	211.16926	C13H22O2	3	3E,5E-tridecadienoic acid [unsaturated fatty acids [FA0103]] ([M+H]+)
E8	213.19614	C12H24N2O	2	0
E8	216.12304	C10H17NO4	3	2-amino-9,10-epoxy-8-oxodecanoic acid ([M+H]+)
E8	222.13494	C10H15N5O	6	dihydrozeatin
E8	223.06347	C8H14O5S	4	2-(3'-methylthio)propylmalic acid
E8	225.03449	C7H13O4SP	5	0
E8	225.12337	C11H16N2O3	5	0
E8	230.24784	C14H31NO	0	xestoaminol C
E8	231.06282	C6H15O7P	1	0
E8	256.14203	C13H19N3O	6	0
E8	235.11762	C10H18O6	2	0
E8	235.20564	C16H26O	4	4,6,11-hexadecatrienal [fatty aldehydes [FA06]] ([M+H]+)
E8	235.98108	C3H10NO5S2P	5	0
E8	258.11861	C8H13N9	7	0
E8	236.20089	C15H25NO	4	0
E8	238.12027	C9H20NO4P	2	0
E8	240.20704	C13H25N3O	3	0
E8	258.27914	C16H35NO	0	0
E8	282.01717	C6H14NO6SP	4	0
E8	263.07364	C14H14O3S	10	0
E8	265.13117	C10H21N2O4P	3	0
E8	293.20872	C16H30O3	2	3-oxohexadecanoic acid
E8	294.20397	C15H29NO3	2	tridecanoylglycine [carboxylic acid] ([M+H]+)
E8	272.99324	C7H14O3S2P2	8	0

E8	278.17507	C16H23NO3	6	0
E8	281.24751	C18H32O2	3	linoleate
E8	285.21727	C15H28N2O3	3	0
E8	287.00889	C8H16O3S2P2	8	0
E8	293.08421	C15H16O4S	10	0
E8	294.15473	C12H23NO7	2	0
E8	318.24035	C18H33NO2	3	(4E,8E,10E-d18:3)sphingosine
E8	300.25332	C17H33NO3	2	pentadecanoylglycine [carboxylic acid] ([M+H]+)
E8	302.15982	C14H23NO6	4	0
E8	306.20637	C18H27NO3	6	capsaicin ([M+H]+)
E8	307.09412	C12H19O7P	5	0
E8	315.05417	C14H18O2S3	12	0
E8	339.17646	C13H20N10	9	0
E8	322.2588	C16H35NO5	0	0
E8	326.37813	C22H47N	0	0
E8	331.19039	C20H26O4	8	carnosol ([M+H]+)
E8	338.25372	C16H35NO6	0	0
E8	361.19855	C19H30O5	5	shiromodiol diacetate ([M+H]+)
E8	349.27372	C22H36O3	5	anacardic acid ([M+H]+)
E8	350.2901	C18H39NO5	0	0
E8	351.0992	C17H19O6P	10	0
E8	376.14005	C14H27NO7S	4	0
E8	377.19346	C19H30O6	5	5-hydroperoxy-7-[3,5-epidioxy-2-(2-octenyl)-cyclopentyl]-6-heptenoic acid [hydroperoxy fatty acids [FA0104]] ([M+H]+)
E8	356.20677	C18H29NO6	5	0
E8	360.15792	C17H30NOS2P	9	0
E8	363.16246	C20H26O4S	10	0

E8	365.10534	C18H20O6S	11	0
E8	387.2142	C21H32O5	6	urocortisone ([M+H]+)
E8	371.09728	C16H18O10	8	fraxin ([M+H]+)
E8	378.21524	C16H32N3O5P	4	0
E8	383.33084	C27H42O	7	7-dehydrodesmosterol ([M+H]+)
E8	388.21185	C22H29NO5	9	0
E8	396.2592	C18H37NO8	1	0
E8	400.38975	C23H49N3O2	1	0
E8	427.06781	C17H24O5S3	12	0
E8	406.24355	C19H35NO8	3	0
E8	455.34688	C24H44N6O	6	0
E8	436.30912	C22H45NO5S	3	0
E8	441.32961	C20H40N8O3	5	0
E8	445.28814	C18H36N8O5	5	0
E8	453.25892	C19H44N4P4	6	0
E8	454.38909	C27H51NO4	3	0
E8	463.10385	C18H28N2O4S2P2	13	0
E8	465.12726	C15H28O14S	4	0
E8	477.3323	C27H44N2O5	7	0
E8	488.26428	C27H37NO7	10	0
E8	488.39457	C27H53NO6	2	0
E8	493.34968	C25H44N6O4	7	0
E8	559.13129	C21H25N6O9P	14	0
E8	539.16602	C27H26N2O10	16	0
E8	554.17589	C21H28N7O9P	13	0
E8	583.20684	C25H34N4O1	13	0

		0S		
E8	629.19381	C22H43N2O1 1S2P	8	0

8 Bibliography

Amann, A., Miekisch, W., Pleil, J., Risby, T., Schubert, J. (2010). Chapter 7: Methodological issues of sample collection and analysis of exhaled breath. *European Respiratory Monograph* 49, 96–114.

Bales, J.R., Higham, D. P., Howe, I., Nicholson, J. K and Sadler, P. J. (1984). Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. *Clinical Chemistry* 30, 426-432.

Barabási, A-L., Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5, 101-113.

Bertini, I., Luchinat, C., Miniati, M., Monti, S., Leonardi, T. (2013). Phenotyping COPD by ^{1}H NMR metabolomics of exhaled breath condensate. *Metabolomics*, 10.1007/s11306-013-0572-3.

Boutegrabet, L., Kanawati, B., Gebefügi, I., Peyron, D., Cayot, P., Gougeon, R. D., Schmitt-Kopplin, Ph. (2012). Attachment of chloride anion to sugars: mechanistic investigation and discovery of a new dopant for efficient sugar ionization/detection in mass spectrometers. *Chemistry – A European Journal* 18, 13059-13067.

Breimann, L., (2001). Random Forests. *Machine Learning 45*, 5-32.

Breitling, R., Shawn, R., Goodenowe, Dayan., Stewart, M. L., Barrett, M. P. (2006). Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics* 2, 155-164.

Byrdwell, Wm. C. (1998). Dual Parallel Mass Spectrometers for Analysis of Sphingolipid, Glycerophospholipid and Plasmalogen Molecular Species. *Rapid Communications in Mass Spectrometry* 12, 256-272.

Campos, A. C. E., Molognoni, F., Melo, F. H. M., Galdieri, L. C., Carneiro, C. R. W., D'Almeida, V., Correa, M., Jasiulionis, M. G. (2007). Oxidative Stress Modulates DNA Methylation during Melanocyte Anchorage Blockade Associated with Malignant

Transformation. *Neoplasia* 9, 1111-1121.

Cao, W. and Duan, Y. (2006). Breath Analysis: Potential for Clinical Diagnosis and Exposure Assessment. *Clinical Chemistry* 52, 800-811.

Chech, N. B., and enke, C. G. (2001). Practical Implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews* 20, 362-387.

Chen, S., Zieve, L., Mahadevan, V. (1970). Mercaptans and dimethylsulfide in the breath of patients with cirrhosis of the liver. *Journal of Laboratory and Clinical Medicine* 75, 628–635.

Cheng, S., Rhee, E. P., Larson, M. G., Lewis, G. D., McCabe, E. L., Shen, D., Palma, M. J., Roberts, L. D., Dejam, A., Souza A. L., Deik, A. A., Magnusson, M., Fox, C. S., O'Donnell, C. J., Vasan, R. S., Melander, O., Clish, C. B., Gerszten, R. E., and Wang, T. J. (2012). Metabolite Profiling Identifies Pathways Associated With Metabolic Risk in Humans. *Circulation* 125, 2222-2231.

Cleveland, W.S., (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association 74*, 829-836.

Cole, R. B., and Harrata, A. K., (1993). solvent effect on analyte charge state, signal intensity, and stability in negative ion electrospray mass spectromentry; implications for the mechanism of negative ion formation. *Journal of the American Society for Mass Spectrometry* 4, 546-556.

Cortes, C., Vapnik, V., (1995). Support-vector networks. *Machine Learning 20*, 273-297.

Dannecker Jr., J. R., Shaskan E. G., Phillips, M. (1981). A new highly sensitive assay for breath acetaldehyde: Detection of endogenous levels in humans. *Analytical Biochemistry* 114, 1-7.

Davis, M. D., Montpetit, A., Hunt, J. (2012). Exhaled Breath Condensate: An Overview. *Journal of Immunology and Allergy Clinics of North America* 32, 363-375

Davison, P.G. (2004). How to Define Life. *The University of North Alabama*. URL: http://www.una.edu/faculty/pgdavison/BI%20101/Overview%20Fall%202004.htm

De Souza, A. G., MacCormack, T. J., Wang, N., Li, L., and Goss, G. G. (2009). Large Scale Proteome Profile of the Zebrafish (*Danio rerio*) Gill for Physiological and Biomarker Discovery Studies. *Zebrafish* 3, 229-238.

Do, R., Bartlett, K. H., Chu, W., Dimich-Ward, H., Kennedy, S. M. (2008). Within- and between-person variability of exhaled breath condensate pH and NH_4^+ in never and current smokers. *Respiratory Medicine* 102, 457-463.

Dwyer, T. M. (2004). Sampling Airway Surface Liquid: Non-Volatiles in Exhlaed Breath Condensate. *Lung* 182, 241-250.

Effros, R. M., Hoagland K. W., Bosbous, M., Castillo, D., Foss, B., Dunning, M., Gare, M., Lin, W. and Sun, F. (2002). Dilution of respiratory solutes in exhaled breath condensate. *American Journal of Respiratory and Critical Care Medicine* 165, 663-669.

Effros, R. M. (2010). Exhaled Breath Condensate: Delusion or Dilution? *Chest* 138, 471-472.

Eliot, A. C., Kirsch, J. F. (2004). Pyridoxal Phosphate Enzymes: Mechanistic, Structural, and Evolutionary Considerations. Annual Review of Biochemistry 73, 383-415.

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology* 48, 155-171.

Flerus, R., Lechtenfeld, O.J., Koch, B.P., MacCallister, S.L., Schmitt-Kopplin, P., Benner, R., Kaiser, K., Kattner, G. (2012). A molecular perspective on the ageing of marine dissolved organic matter. *Biogeosciences 9*, 1935-1955.

Forcisi, S., Moritz, F., Kanawati, B., Tziotis, T., Lehmann, R., and Schmitt-Kopplin, Ph. (2013). Liquid chromatography-mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of Chromatography A* 1292, 51-65.

Garcia-Ac, A., Segura, P. A., Viglino L., Gagnon, C., Sauvé, S. (2011). Comparison of APPI, APCI, and ESI for the LC-MS/MS analysis of bezafibrate, cyclophosphamide, enalapril, methotrexate and orlistat in municipal wastewater. *Journal of Mass Spectrometry* 46, 383-390.

Gartland, K.P.R., Bonner, F. W., Nicholson, J. K. (1989). Investigations intothe biochemical effects of region-specific nephrotoxins. *Molecular Pharmacology* 35, 242-250.

Gavaghan, C. L., Holmes, E., Lenz, E., Wilson, I. D., Nicholson, J. K. (2000). An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: applicaiton to the C57BL10J and Alpk:ApfCD mouse. *FEBS Letters* 484, 169-174.

Girvan, M., Newman, M. E. J. (2002). Community structure in social and biological networks. *PNAS* 99, 7821-7826.

Goldman, M. D. (2003). Lung Dysfunction in Diabetes. *Diabetes Care* 26, 1915-1918.

Guihaus, M. (1995). Principles and Instrumentation in Time-of-flight Mass Spectrometry. *Journal of Mass Spectrometry* 30, 1519-1532.

Guimerà, R., Nunes Amaral, L. A. (2005). Functional cartography of complex metabolic networks. *Nature* 433, 895-900.

Haenlein M. and Kaplan, A. M. (2004). Volume 3, Issue 4: “Understanding Statistics” in A Beginner’s Guide to Partial Least Squares Analysis, 283-297.

Harris, M., and Zimmet, P. (1997). Classification of diabetes mellitus and other categories of glucose intolerance. In Alberti, K., Zimmet, P., Defronzo, R., editors. *International Textbook of Diabetes Mellitus*. Second Edition. Chichester: John Wiley and Sons Ltd; p9-23.

Henriksen T., Juhler, R. K., Scensmark, B., Cech, N. B. (2005). The relative influences of acidity and polarity on responsiveness of small organic molecules to analysis with negative ion electrospray ionization mass spectrometry (ESI-MS). *Journal of the American Society for*

Mass Spectrometry 16, 446-455.

Hertkorn, N., Frommberger, M., Witt, M., Koch, B. P., Schmitt-Kopplin, Ph., Perdue, E. M. (2008). Natural Organci Matter and the Event Horizon of Mass Spectrometry. *Analytical Chemistry* 80, 8908-8919.

Holmes, E., Wilson, I. D., Nicholson, J. K. (2008). Metabolic Phenotyping in Health and Disease. *Cell* 134, 714-717.

Horvath, I., Hunt, J., Barnes, P. J. (2005). Exhaled breath condensate: methodological recommendations and unresolved questions. *European Respiratory Journal* 26, 523-548.

Huffmann, K.M., Shah, S. H., Stevens, R. D., Bain, J. R., Muehlbauer, M., Slentz, C. A., Tanner, C. J., Kuchibhatla, M., Houmard, J. A., Newgard, C. B., Kraus, W. E., (2009). Relationships Between Circulating Metabolic Intermediates and Insulin Action in Overweight to Obese, Inactive Men and Women. *Diabetes Care* 32, 1678-1683.

Hunt, J. F. (2002). Exhaled Breath Condensate: An evolving tool for non.invasive evaluation of lung disease. *Journal of Allergy and Clinical Immunology* 110, 28-34.

International Diabetes Federation. IDF Diabetes Atlas, 5th edn. Brussels, Belgium: International Diabetes Federation, 2011.

Jansson, B. O., Larsson, B. T. (1969). Analysis of organic compounds in human breath by gas chromatography-mass spectrometry. *Journal of Laboratory and Clinical Medicine* 74, 961-966.

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., Schmitt-Kopplin, Ph. (2009). Metabolomics Reveals Metabolic Biomarkers of Crohn's Disease. PLoS ONE 4(7): e6386. doi:10.1371/journal.pone.0006386.

Kedrick, E. (1963). A Mass Scale Based on CH_2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Anal. Chem. 35*, 2146-2154.

Kharitonov, S. A., Barnes, P. J. (2001). Exhaled Markers of Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* 163, 1693-1722.

Kind, T., Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* 8, 105.

Klitzke, C. F., Corilo, Y. E., Siek, K., Binkley, J., Patrick, J., Eberlin, M. N. (2012). Petroleomics by Ultrahigh-Resolution Time-of-Flight Mass Spectrometry. *Energy Fuels* 26, 5787-5794.

Koch, B.P., Dittmar, T. (2006). From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry 20*, 926-932.

Krug, S., Kastenmüller, G., Stückler, F., Rist, M. J., Skurk, T., Sailer, M., Raffler, J., Römisch-Margl, W., Adamski, J., Prehn, C., Frank, T., Engel, K-H., Hofmann, T., Luy, B., Zimmernmann, R., Moritz, F., Schmitt-Kopplin, Ph., Krumsiek, J., Kremer, W., Huber, F., Oeh, U., Theis, F. J., Szymczak, W., Hauner, H., Suhre, K., Daniel, H., (2012). The dynamic range of the human metabolome revealed by challenges. *The FASEB Journal* 26, 2607-2619.

Kumar, M., Sarin, S. K. (2009). Biomarkers of diseases in medicine. *Current Trends in Science*, Platinum Jubilee Special, 403-417.

Li, L., Alderson, D., Tanaka, R., Doyle, J. C., Willinger, W. (2005). Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. Cornell University Library, arXiv:cond-mat/0501169v2.

Li, M., Wang, B., Zhang, M., Rantalainen, M., Wang, S., Zhou, H., Zhang, Y., Shen J., Pang, X., Zhang, M., Wei, H., Chen, Y., Lu, H., Zuo, J., Su, M., Qiu, Y., Jia, W., Xiao, C., Smith, L. M., Jolmes, E., Tang, H., Zhao, G., Nicholson, J. K., Li, L. Zhao, L. (2008) Symbiotic gut microbes modulate human metabolic phenotypes. *PNAS* 105, 2117-2122.

Lindon J. C., Nicholson, J. K. (2008). Spectroscopic and Statistical Techniques for Information Recovery in Metabonomics and Metabolomics. *Annual Review of Analytical*

Chemistry 1, 45-69.

Lu., T., Costello, C. M., Croucher, P. J. P., Häsler, R., Deuschl, G., Schreiber, S. (2005). Can Zipf's law be adapted to normalize microarrays? *BMC Bioinformatics* 6:37, doi:10.1186/1471-2105-6-37.

Mamyrin, B.A. (2001). Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry* 206, 251-266.

Marshall A., Hendrickson, C. L., Jackson, G. S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews* 17, 1-35.

Matsuda, M., DeFronzo, R. A. (1999). Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care* 22, 1462-1470.

McKay, C. P. (2004) What Is Life—and How Do We Search for It in Other Worlds? *PLoS Biology* 2(9), doi:10.1371/journal.pbio.0020302.

McPherson, K., Healy, M. J. R., Flynn, F. V., Piper, K. A. J., Garcia-Webb, P. (1978). The effect of age, sex and other factors on blood chemistry in health. Clinica Chimica Acta 84, 373-397.

Medvei, V.C. (1993) The History of Clinical Endocrinology: A Comprehensive Account of Endocrinology from Earliest Times to the present Day. ISBN-10: 1850704279; ISBN-13: 978-1850704270.

Montuschi, P., Paris, D., Melck, S., Lucidi, V., Ciabattoni, G., Raia, V., Calabrese, C., Bush, A., Barnes, P. J., Motta, A. (2012). NMR spectroscopy metabolomics profiling of exhaled breath condensate in patients with stable and unstable cystic fibrosis. *Thorax* 67, 222-228.

Moka D., Vorreuther, R., Schicha, H., Spraul, M., Humpfer, E., Lipinski, M., Foxall, P. J. D., Nicholson, J. K., Lindon, J. C. (1998). Biochemical classification of kidney carcinoma biopsy samples using magic.angle.spinning ^{1}H nuclear resonance spectronscopy. *Journal of*

Pharmaceutical and Biomedical Analysis 17, 125-132.

Möller, W., Heimbeck, I., Weber, N., Khadem Saba, G., Körner, B., Neiswirth, M., Kohlhäufl, M. (2010). Fractionated Exhaled Breath Condensate Collection Shows High Hydrogen Peroxide Release in the Airways. *Journal of Aerosol Medicine and Pulmonary Drug Delivery* 23, 129-135.

Morikawa, T., Newbold, B. T., (2003). Analogous odd-even parities in mathematics and chemistry. *Chemistry* 12, 445-450.

Mutlu, G. M., Garey, K. W., Robbins, R. A., Danziger, L. H., Rubinstein, I. (2001). Collection and Analysis of Exhaled Breath Condensate in Humans. *American Journal of Respiratory and Critical Care Medicine* 164, 731-737.

Nagarajan, N., Pop, M. (2010). Sequencing and Genome Assembly Using Next-Generation Technologies. *Computational Biology - Methods in Molecular Biology* 673, 1-17.

Newman, M. E. J. (2004a). Fast algorithm for detecting community structure in networks. *Physical review E* 69, DOI: 10.1103/PhysRevE.69.066133.

Newman, M. E. J., Girvan, M. (2004b). Finding and evaluating community structure in networks. *Physical review E* 69, DOI:10.1103/PhysRevE.69.026113.

Nicholson J. K., Buckingham, M. J., Sadler, P. J. (1983). High Resolution 1H n.m.r. studies of vertebrate blood and plasma. Biochemical Journal 211, 605-615.

Nicholson J. K., Holmes, E., Wilson, I. D. (2005). Gut microorganisms, mammalian metabolism and personalized health care. *Nature Reviews Microbiology* 3, 431-438.

Nicholson J. K., Connelly, J., Lindon, J. C., Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery* 1, 153-161.

Nicholson J. K., Wilson, I. D. (1989). High resolution proton magnetic resonance spectroscopy of biological fluids. *Progress in Nuclear Magnetic Resonance Spectroscopy* 21, 449-501.

Nicholson, J. K., Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature* 456, 1054-1056.

Nicholson, J.K., Timbrell, J. A., Sadler, P. J. (1985). Proton NMR spectra of urine as indicators of renal damage. Mercury-induced nephrotoxicity in rats. Molecular Pharmacology 27, 644-651.

Nicholson, J.K., Lindon, J. C., Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1181-1189.

Nicholson, J.K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular Systems Biology* 2, 1-6.

Oss, M., Kruve, A., Herodes, K., Leito, I. (2010). Electrospray Ionization Efficiency Scale of Organic Compounds. *Analytical Chemistry* 82, 2865-2872.

Pauling, L., Robinson, A. B., Teranishi, R., Cary, P. (1979). Quantitative Analysis of Urine Vapor and Breath by Gas-Liquid Partition Chromatography. *PNAS* 68, 2374-2376.

Pellegrin, V. (1983). Molecular Formulas of Organic Compounds. *Journal of Chemical Education 60*, 626-633.

Phillips, M., Herrera, J., Krishnan, S., Zain, M., Greenberg, J., Cataneo, R. N. (1999). Variation in volatile organic compounds in the breath of normal humans. *Journal of Chromatography B: Biomedical Sciences and Applications* 729, 75-88.

Phillips, M. (1992). Breath Tests in Medicine. *Scientific American* 267, 74-79.

Rennard, S. I., Basset, G., Lecossier, D., O'Donnell, K. M., Pinkston, P., Martin, P. G., Crystal, R. G. (1986) Estimation of Volume of epithelial lining fluid recovered by lavage using urea as marker of dilution. *Journal of Applied Physiology* 60, 532-538.

Riely, C. A., Cohen, G., Liebermann, M. (1974). Ethane Evolution: A New Index of Lipid Peroxidation. *Science* 183, 208-210.

Risby, T. H., Sehnert, S. S. (1999). Clinical application of breath biomarkers of oxidative stress status. *Free Radical Biology and Medicine* 27, 1182-1192.

Risby, T. H., Solga, S. F. (2006). Current status of clinical breath analysis. *Applied Physics B* 85, 421-426.

Risby, T. H. (2002). Volatile organic compounds as markers in normal and diseased states. In: Marczin, N. and Yacoub, M. H., (Eds.) Disease Markers in Exhaled Breath: Basic Mechanisms and Clinical Applications. NATO ASI Series, IOS Press, Amsterdam, pp. 113-122.

Savory, J. J., Kaiser, N. K., McKenna, A. M., Xian, F., Blakney, G. T., Rodgers, R. P., Hendrickson, C. L., Marshall, A. G. (2011). Parts-Per-Billion Fourier Transform Ion Cyclotron Resonance Mass Measurement Accuracy with a “Walking” Calibration Equation. *Analytical Chemistry* 83, 1732-1736.

Schmitt-Kopplin. Ph., Gabelica, Z., Gougeon, R. D., Fekete, A., Kanawati, B., Harir, M., Gebefuegi, I., Eckel, G., Hertkorn, N. (2010). High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *PNAS* 107, 2763-2768.

Senior, J. K. (1951). Partitions and their representative graphs. *American Journal of Mathematics* 73, 663-689.

Sofia, M., Maniscalco, M., de Laurentiis, G., Paris, D., Melck, D., Motta, A. (2011). Exploring Airway Diseases by NMR-Based Metabonomics: a Review of Application to Exhaled Breath Condensate. *Journal of Biomedicine and Biotechnology* (2011) Article ID 403260.

Soininen P., Kangas, A. J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., Järvelin, M-R., Kähönen, M., Lehtimäki, T., Viikari, J., Raitakari, O. T., Savolainen, M. J., Ala-Korpela, M. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* 134, 1781-1785.

Solga, S. F., Risby, T. H. (2010). What is Normal Breath? Challenge and Opportunity. *IEEE Sensors Journal* 10, 7-9.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* 102, 15545-15550.

Schmidt, G.A. (2010). Counterpoint: adherence to early Goal-Directed Therapy: Does it Really Matter? No. Both Risks and Benefits Require further Study. *Chest* 138, Editorial.

Suhre, K., Meisinger, M., Döring, A., Altmaier, E., Belcredi, P., Gieger, C., Chang, D., Milburn, M. V., Gall, W. E., Weinberger, K. M., Mewes, H-W., Hrabé de Angelis, M., Wichmann, H-E., Kronenberg, F., Adamski, J., Illig, T. (2010). Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. PLoS ONE 5, doi:10.1371/journal.pone.0013953.

Suhre. K., Schmitt-Kopplin, Ph. (2008). MassTRIX: mass translator into pathways. *Nucleic Acids Research* 36, 481-484.

Tziotis, D., Hertkorn, N., Schmitt-Kopplin, Ph. (2011). Kendrick-analgous network visualization of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *European Journal of Mass Spectrometry* 17, 415-421.

Van den Berg, R.A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7, 142-158.

Villas-Boas, S.G., et al., Metabolome Analysis: An Introduction. 2007, Wiley.

Von der Malsburg, Chr. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik 14*, 85-100.

Wang, T.J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., Lewis, G. D., Fox, C. S., Jacques, P. F., Fernandez, C., O'Donnell, C. J., Carr, S. A., Mootha, V. K., Florez, J. C., Souza, A., Melander, O., Clish, C. B., Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine* 17, 448-453.

Willinger, W., Alderson, D, Doyle, J. C., Li, L. (2004). More „normal" than Normal: Scaling distributions and complex systems. *Proceedings of the 2004 Winter Simulation Conference*.

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems 58*, 109-130.

Würtz, P., Mäkinen, V-P., Soinen, P., Kangas, A. J., Tukiainen, T., Kettunen, J., Savolainen, M. J., Tammelin, T., Viikari, J. S., Rönnemaa, T., Kähönen, M., Lehtimäki, T., ripatti, S., Raitakari, O. T., Järvelin, M-R., Ala-Korpela, M. (2012). Metabolic Signatures of Insulin Resistance in 7,098 Young Adults. *Diabetes* 61, 1372-1380.

9 Acknowledgements

I want to thank Prof. Philippe Schmitt-Kopplin for introducing me into this incomparable group with its incomparable people and its incomparable technical setup. I want to thank him for always supporting and believing in me and for providing the scientific freedom that is necessary for creative research.

I want to thank Dr. Istvan Gebefügi for the hours of philosphy, encouragement, friendliness and for sharing the knowledge that he had gained throughout his professional career.

I want to thank Prof. Hannelore Daniel and Dr. Susanne Krug for organizing and performing the HuMet study and to enable us to perform our analyses on this very valuable sample set.

I want to thank all my collegues for their personal and professional intercourse. It was really a pleasant time being with you!

Special thanks to Sara for simply being who she is and for showing as much dedication to my problems and life as I have shown for hers. We have shared very much and being with you was a very important time.

Special thanks to my best friend and eternal best office-mate Dimitris. Thank you for sharing all ups and downs, for friendship beyond work, for what you have tought me – simply: thank you!

In addition, I want to thank Stefan Bardtke and Julia Bardtke for their unconditional human and technical support and for simply being friends.

Und schließlich, ein unbegrenzbar großer Dank an meine Familie! An meine Mutter – Josta – und meinen Vater – Falko – die mich seit ich denken kann mit keinem Schimmer des Zweifelns unterstützt haben; die immer an mich geglaubt haben und immer hinter mir standen. An meine Geschwister Caroline, Karl und Konrad; dafür dass ihr das Leben lebenswert macht. Caro, ich bewundere deinen Aktionismus! Ich kann noch einiges von dir lernen. Karl und Konrad, ihr seid einfach super. Ich hab euch beide lieb. An Oma Moni, welche die vielleicht beste Großmutter ist, die man sich wünschen kann. An Opa Peter, der durch stetige Herausforderung aber auch offene (manchmal forsche) Fürsorge unterstützt hat. Und auch an Oma Margot, welcher unter die Augen zu treten ich mich seit mittlerweile 10 Jahren nicht getraut habe – trotz ihrer Liebenswürdigkeit. Ich danke auch Onkel Otmar und Tante Anja; ihr seid einfach toll. Und nicht zu vergessen Oma Rosi und Opa Bernd, die uns – meine Schwester und mich – unvoreingenommen als Enkel angenommen haben.

10 Scientific Communications

- Franco Moritz, Sara Forcisi, Mourad Harir, Basem Kanawati, Marianna Lucio, Dimitrios Tziotis, Philippe Schmitt-Kopplin
  Book chapter;
  The potential of ultrahigh resolution MS (FTICR-MS) in metabolomics
  Edited by Michael Lämmerhofer and Wolfram Weckwerth
  ISBN: 978-3-527-33089-8

- Susanne Krug, Gabi Kastenmüller, Ferdinand Stückler, Manuela J. Rist, Thomas Skurk, Manuela Sailer, Johannes Raffler, Werner Römisch-Margl, Jerzy Adamski, Cornelia Prehn, Thomas Frank, Karl-Heinz Engel, Thomas Hoffmann, Burkhard Luy, Ralf Zimmermann, Franco Moritz, Philippe Schmitt-Kopplin, Jan Krumsiek, Werner Kremer, Fritz Huber, Uwe Oeh, Fabian J. Theis, Wilfried Szymczak, Hans Hauner, Karsten Suhre and Hannerlore Daniel.
  The dynamic range of the human metabolome revealed by challenges.
  Faseb (2012), doi: 10.1096/fj.11-198093

- Constanze Müller, Inga Dietz, Dimitrios Tziotis, Franco Moritz, Jan Rupp, Philipper Schmitt-Kopplin
  Molecular cartography in acute *Chlamydia pneumoniae* infections – an non-targeted metabolomics approach.
  Anal Bioanal Chem, doi: 10.1007/s00216-013-6732-5

- Josefa Antón, Marianna Lucio, Ana Cifuentes, Jocely Brito-Echeverria, Franco Moritz, Dimitrios Tziotis, Cristina López, Mercedes Urdiain, Philippe Schmitt-Kopplin, Ramon Rosselló-Móra
  High Metabolomic Microdiversity within Co-Occurring Isolates of the Extremely Halophilic Bacterium *Salinibacter ruber*.
  PloS One, doi:10.1371/journal.pone.0064701

- Sara Forcisi, Franco Moritz, Basem Kanawati, Dimitrios Tziotis, Rainer Lehmann, Philippe Schmitt-Kopplin
  Liquid chromatography-mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling.
  Journal of Chromatography A, Volume 1292, 31 May 2013, Pages 51–65

- Kilian Wörmann, Alesia Walker, Franco Moritz, Sara Forcisi, Dimitrios Tziotis, Marianna Lucio, Silke Heinzmann, Jerzy Adamski, Rainer Lehmann, Hans-Ulrich Häring, Philippe Schmitt-Kopplin
  Revolution in Diabetes Diagnostics – Metabolomics for Discovering Biomarkers.
  Diabetes aktuell für die Hausarztpraxis, ISSN 1864-1733

11 Curriculum vitae

Franco Moritz

Persönliche Informationen
Alter: 28 Jahre
Gerburtsdatum: 26. September 1984 (Dresden, Deutschland)
Nationalität: deutch
Email: franco.moritz@gmail.com

BILDUNG

Oktober 2009 – Mai 2013
Doktorand am Helmholtz Zentrum München
Arbeitstitel: Deep Metabotyping of exhaled breath condensate (EBC) – characterization of surrogate markers for systemic metabolism and non-invasive diagnostics in Diabetes.
Betreuung: apl Prof. Dr. Philippe Schmitte-Kopplin (Technische Universität München, Lehrstuhl für Analytische Lebensmittelchemie, Alte Akademie, Freising, Deutschland)

17. September 2009
Titel der erworbenen Qualifikation:
Abschluss zweiten Grades (Msc) in Umwelt-Biotechnologie und angewandter Ökologie (Spezialisierung: Umwelt)
Arbeitstitel: Metabolomvergleich von Atemkondensaten und Blutplasma mittels hochauflösender Fourier Transform Ionencyclotron Resonanz

Massenspektrometrie (FT-ICR-MS)

Betreuung: Priv. Doz. Dr. Ph. Schmitt-Kopplin (WZW), Dr. rer. Nat habil. Stefan Fränzle (IHI Zittau)

Name der Bildungseinrichtung: Internationales Hochschulinstitut Zittau.

19. Dezember 2007

Titel der erworbenen Qualifikation:

Abschluss ersten Grades (BSc) in Biotechnologie (Spezialisierung: Verfahrenstechnik)

Arbeitstitel: Reinigunf und partielle Characterisierung einer extrazellulären Esterase des Acomyzeten *Xylaria polymorpha*.

Betreuung: Prof. Fuchs (Professorin für Biochemie FH Zittau-Görlitz)

Name der Bildungseinrichtung: Fachhochschule Zittau-Görlitz

ARBEITSERFAHRUNG

Oktober 2009 – jetzt:

(IÖC, jetzt BGA, TUM), ICR-FT-MS Profiling und Datenanalyse von Atemkondensat, Blutplasma, Urin, Fäzes, Pflanzenextrakten sowie Bakterienkulturen. Entwicklung Datenanalytischer Methoden.

TRAINING UND KURSE

11.-12. Oktober 2011:

Wissenschaftliches Management, Training, Workshop (ReMaT)

Gehalten von TuTech Innovation, Helmholtz Assoziation von Deutschen Wissenschaftszentren, Abteilung Hamburg und Brüssel

FÄHIGKEITEN UND KOMPETENZ

Sprache:

Deutsch: Muttersprache

Englisch: flüssig in Schrift und Sprache

Italienisch: Grundstufe

Technische Fähigkeiten und Kompetenzen:

Probenaufbereitung und Manipulation von biologischen Proben, chemische Analyse mittels FT-ICR-MS, UPLC-TOF-MS, FPLC, Verschieden Arten der Elektrophorese (präp-IEF, SDS-PAGE, im weitesten sinne on-column IEF) Multivariate Analyse, Graphentheorie, Matlab sowie Interpretation biologischer Daten

Computer-Fähigkeiten und Kompetenzen:

Betriebssystem: Windows

Anwendungen: MS Office (Word, Excel, Powerpoint), Gephi, Matlab, Mathematica

Software zur Datenanalyse: Data Analysis (Bruker), mzmine, Datenbanken zur Massenannotierung (Pubchem, MassTRIX, ChemSpider),

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Promotionsprüfung vorgelegte Arbeit mit dem Titel: „Deep Metabotyping of exhaled breath condensate (EBC) – characterization of surrogate markers for systemic metabolism and non-invasive diagnostics in Diabetes“ unter der Anleitung und Betreuung durch apl. Prof. Dr. Philippe Schmitt-Kopplin ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß §6 Abs. 5 angegebenen Hilfsmittel benutzt habe.

(x) Ich habe die Dissertation in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt.

(x) Ich habe den angestrebten Doktorgrad noch nicht erworben und bin nicht in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert.
Die Promotionsordnung der Technischen Universität München ist mir bekannt.

München, den 18.09.2013