



PhD-FSTC-2013-35

The Faculty of Sciences

DISSERTATION

Presented on 12/12/2013 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

by

Andre WEGNER

Born on 28 June 1982 in Berlin (Germany)

COMPUTATIONAL TOOLS FOR MASS SPECTROMETRY
BASED METABOLIC PROFILING

Dissertation defense committee

Dr Karsten Hiller, dissertation supervisor
Université du Luxembourg

Dr Rudi Balling, Chairman
Professor, Université du Luxembourg

Dr Reinhard Schneider Vice Chairman
Université du Luxembourg

Dr Royston Goodacre
Professor, University of Manchester

Abstract

Within systems biology metabolomics emerged as an important field to study cellular metabolism. The successful application of metabolomics techniques requires a close interplay between experimental and computational approaches. In particular, stable isotope assisted metabolomics methodologies require specialized algorithms to extract biological information out of a metabolomics experiments. In this thesis the development and application of novel mass spectrometry-based algorithms to analyze cellular metabolism are presented.

First, the isotope cluster based matching (ICBM) algorithm was developed and implemented as a c++ based library. The ICBM algorithm is a spectrum similarity measure that is most efficient for the matching of compounds across different chromatograms. Especially for non-targeted analyses, the ICBM algorithm outperforms the dot product and other conventional tools. Moreover, the ICBM algorithm can be applied for an efficient mass spectral library search.

Second, the ICBM algorithm was applied to characterize the metabolomes of the human neuronal cell line LUHMES at low oxygen conditions (5%) compared to standard cell culture conditions (20%). A difference at the metabolite level was observed when cells were differentiated at 5% oxygen. Beside others, the major inhibitory neurotransmitter in the mammalian central nervous system γ -aminobutyric acid (GABA) was found to be increased at 5% oxygen.

Third, a methodology for the determination of chemical formulas and retained atoms for mass spectral fragment ions was developed. This information about mass spectral fragment ions is indispensable to extract more biological knowledge out of stable isotope labeling experiments. Therefore, the fragment formula calculator (FFC) algorithm was employed to determine the chemical formulas and retained carbon atoms of 160 mass spectral fragment ions of central carbon metabolism.

Acknowledgements

This work is the result of three years of research and there are many people who contributed to it. First, I would like to thank my thesis supervisor Karsten Hiller. This thesis would not have been possible without his guidance, support, and inspiration.

I am very grateful to my parents for giving me the inspiration to work hard and never give up, and for always being there for me when I need them. I am also very grateful to my girlfriend Gintare, who loved and supported me during the whole time of my dissertation, even when I retreated to long days with my computer.

I thank my fellow labmates in the metabolomics groups. In particular Daniel whose attention to detail drove me almost crazy, but made me rethink about problems I would have overlooked without him. I am also grateful to Jenny that she introduced me to working in the cell culture lab. I would also like to thank Sean for proof reading my manuscripts and discussions regarding the english language. Many thanks to Christian who performed all the yeast experiments and to Johannes who performed the TH western blot. Last but not least I would like to thank Nadja for giving me a lift a billion times to Trier.

Further, I would like to thank Gregory Stephanopoulos for giving me the opportunity to stay as visiting student in his lab at MIT.

I would like to thank my defence committee Oliver Kohlbacher, Roy Goodacre, Rudi Balling, and Reinhard Schneider.

Finally, I would like to acknowledge the financial support of the Fond National de la Recherche (FNR).

Contents

Abstract	ii
Acknowledgements	iii
List of Figures	vi
List of Tables	vii
Abbreviations	viii
1 Introduction	1
1.1 Metabolomics	1
1.1.1 Metabolomics Techniques	2
1.1.1.1 Sample Preparation	2
1.1.1.2 Gas Chromatography Coupled To Mass Spectrometry	3
1.1.1.3 Derivatization	6
1.1.2 Methods Of Compound Identification	6
1.1.2.1 Spectrum Similarity Score	7
1.1.2.2 Retention Index Similarity Score	8
1.1.2.3 Chromatogram Alignment	8
1.1.3 Targeted vs Non-Targeted Metabolomics	9
1.2 Stable Isotope Assisted Metabolomics	9
1.2.1 Mass Isotopomer Distribution	10
1.3 Metabolic Flux Analysis	12
1.4 Aim And Outline Of Thesis	18
2 Non-Targeted Metabolomics Methodologies	20
2.1 Isotope Cluster Based Compound Matching	21
2.1.1 Mathematical Description Of The ICBM Algorithm	23
2.1.1.1 Isotope Cluster Determination	24
2.1.1.2 Isotope Cluster Alignment	24
2.1.1.3 Similarity Score Calculation	24
2.1.2 Applications Of The ICBM Algorithm	26
2.2 Non Targeted Tracer Fate Detection	29
2.3 Metabolome Of The Neuronal Cell Line LUHMES	32

2.3.1	Oxygen Level	32
2.3.2	Dopamine Metabolism	34
3	Targeted Metabolomics Methodologies	38
3.1	Fragment Formula Calculator	38
3.1.1	Algorithm	40
3.1.1.1	Reducing Algorithmic Complexity	43
3.1.1.2	Constraining The Result Set	44
3.1.2	Software Package	46
3.2	Fragment Formula Repository	48
3.2.1	TMS Derivatized Fragment Ions	50
3.2.2	TBDMS Derivatized Fragment Ions	53
4	Conclusion	57
	Bibliography	59

List of Figures

1.1	GC/MS scheme	3
1.2	Electron ionization source	4
1.3	Example of rearrangement after EI	5
1.4	GC/MS isotope cluster	6
1.5	Schema of GC/MS derivatization	7
1.6	Mass isotopomer distribution	11
1.7	Simplified scheme of the pentose phosphate pathway and glycolysis	12
1.8	A simple example network	14
1.9	MFA overview	15
1.10	Atom mapping matrix	16
1.11	Value of stable isotope labeling experiments	17
2.1	Example of a spectrum similarity score calculated with the dot product	21
2.2	Overview of the ICBM algorithm	23
2.3	MetaboliteDetector library search	26
2.4	Overview of ICBM library search	28
2.5	NTFD overview	29
2.6	NTFD graphical user interface	30
2.7	LUHMES differentiation	32
2.8	PCA LUHMES	33
2.9	Bar plots LUHMES hypoxia normoxia	35
2.10	GC/MS measurement of mouse midbrain extract	36
2.11	Tyrosine hydroxylase abundance	36
3.1	Importance of targeted methodologies	39
3.2	Overview of FFC algorithm	41
3.3	Graph representation of <i>N,O</i> -bis-(trimethylsilyl)-glycine	43
3.4	Constraining the result set	45
3.5	FFC GUI	47
3.6	FFC library search	48
3.7	Overview yeast culture	49
3.8	Overview yeast measurements	49

List of Tables

3.1	TMS Derivatized Fragment Ions	50
3.2	TBDMS Derivatized Fragment Ions	53

Abbreviations

AMM	A tom M apping M atrice
CE	C apillary E lectrophoresis
CSF	C erebrospinal F luid
EI	E lectron I onization
EMU	E lementary M etabolite U nit
FFC	F ragment F ormula C alculator
GC	G as C hromatography
ICBM	I on C luster B ased M atching
IDV	I sotopomer D istribution V ector
IUPAC	I nternational U nion of P ure and A ppplied C hemistry
LC	L iquid C hromatography
LPS	L ipopolysaccharide
MFA	M etabolic F lux A nalysis
MID	M ass I sotopomer D istribution
MS	M ass S pectrometry
NIST	N ational I nstitute of S tandards and T echnology
NMR	N uclear M agnetic R esonance spectroscopy
NTFD	N on-targeted T racer F ate D etection
PD	P arkinson's D isease
PPP	P entose P hosphate P athway
RI	R etention I ndex
TH	T yrosine H ydroxylase
UPLC	U ltra P erformance L iquid C hromatography

To my parents

Chapter 1

Introduction

This chapter covers the following publication [[Wegner et al., 2012](#)]:

Wegner, A.; Cordes, T.; Michelucci, A.; Hiller, K.
Current Biotechnology (2012), 1, 88-97

1.1 Metabolomics

In the postgenomic era, metabolomics has emerged as an important methodology within systems biology and is defined as the analysis of the set of small molecule (or metabolite) concentrations or amounts produced by a living organism. Although the concepts of metabolomics was grounded more than 40 years ago, the first definition of metabolomics was made by Oliver *et al.* in 1998 [[Oliver et al., 1998](#)]. Certainly, the analysis of the metabolome complements the other three big “omics”, namely genomics, transcriptomics and proteomics, but offers some unique advantages. Since metabolites are the endproduct of the cell’s regulatory processes, the metabolome represents the cell’s final phenotype. Hence it can be considered as the cell’s ultimate response to genetic or environmental perturbations and therefore provides a closer functional link to an observed phenotype [[Villas-Bôas et al., 2005](#)]. In addition, the metabolome could reflect extra-genomic effects caused by factors like the microbiome, which are not accessible by transcriptomics or proteomics [[Hunter, 2009](#)]. As an example, it has been shown that the microbiome has a large effect on the blood metabolome in mice [[Wikoff et al., 2009](#)]. However, the metabolome is much more diverse than the genome or proteome and consists of more than four (in the case of nucleic acids) or twenty (in the case of proteins) unique building blocks. Compared to nucleic acids and proteins, the turnover

rate of metabolites is more than two magnitudes higher and lies in the range of seconds [Sellick et al., 2009a]. Therefore, accurate sampling of the metabolome requires specific techniques, both for the extraction and measurement of metabolites. Current methodologies for quantifying the metabolome typically rely either on nuclear magnetic resonance spectroscopy (NMR) or mass spectrometry (MS) and analyze only the parts of the metabolome which are defined by the methodology [Issaq et al., 2009, Macel et al., 2010].

Generally, these analyses can be distinguished between targeted and non-targeted types. While the first focuses on a set of known metabolites, the second approach tries to get information about all known and unknown detectable metabolites. Although a non-targeted approach provides information about more metabolites and is able to detect changes in unexpected parts of the metabolome, absolute quantitative information can only be obtained by a targeted approach [Hiller et al., 2011].

While metabolomics in its original form quantifies metabolite amounts or concentrations, metabolic flux analysis (MFA) determines absolute values for metabolic conversion rates or fluxes through the metabolic network [Haverkorn van Rijsewijk et al., 2011, Noguchi et al., 2009]. These fluxes are dependent on metabolite concentrations and enzyme activities. Due to the very targeted approach of MFA, its application is limited to known parts of the metabolic network. As an extension, non-targeted stable isotope assisted metabolomics methodologies have been developed. These non-targeted approaches allow to obtain information about the metabolic fate of a labeled compound and can be the starting point for the discovery of unknown or unexpected metabolic pathways [Hiller et al., 2010, Kusmierz and Abramson, 1994, Sano et al., 1976].

1.1.1 Metabolomics Techniques

1.1.1.1 Sample Preparation

A typical metabolomics experiment can be divided in three key steps: sample preparation, analytical metabolite detection and computational data analysis. As the turnover rate of metabolites lies in the range of seconds, a fast and effective quenching procedure is necessary to immediately freeze all biochemical reactions of the cell. To prevent leakage of intracellular metabolites the cell membrane should not be damaged by this process. For this reason, the predominant quenching methods use an ice-cold methanol-water or ethanol water mixture to abolish the tertiary structure of metabolic enzymes, thereby stopping the metabolism [Spura et al., 2009, Villas-Bôas et al., 2005]. A wide range of protocols have been developed for the extraction and quenching of metabolites for mammalian cells grown in suspension [Sellick et al., 2009b], for adherent mammalian

cells, for body fluids (e.g. plasma, urine [Kind et al., 2007], cerebrospinal fluid (CSF) [Wishart et al., 2008]) and for tissue [Wu et al., 2008]. The exact sample preparation protocol not only depends on the biological sample, but also on the targeted metabolites and the analytical technique [Lorenz et al., 2011].

1.1.1.2 Gas Chromatography Coupled To Mass Spectrometry

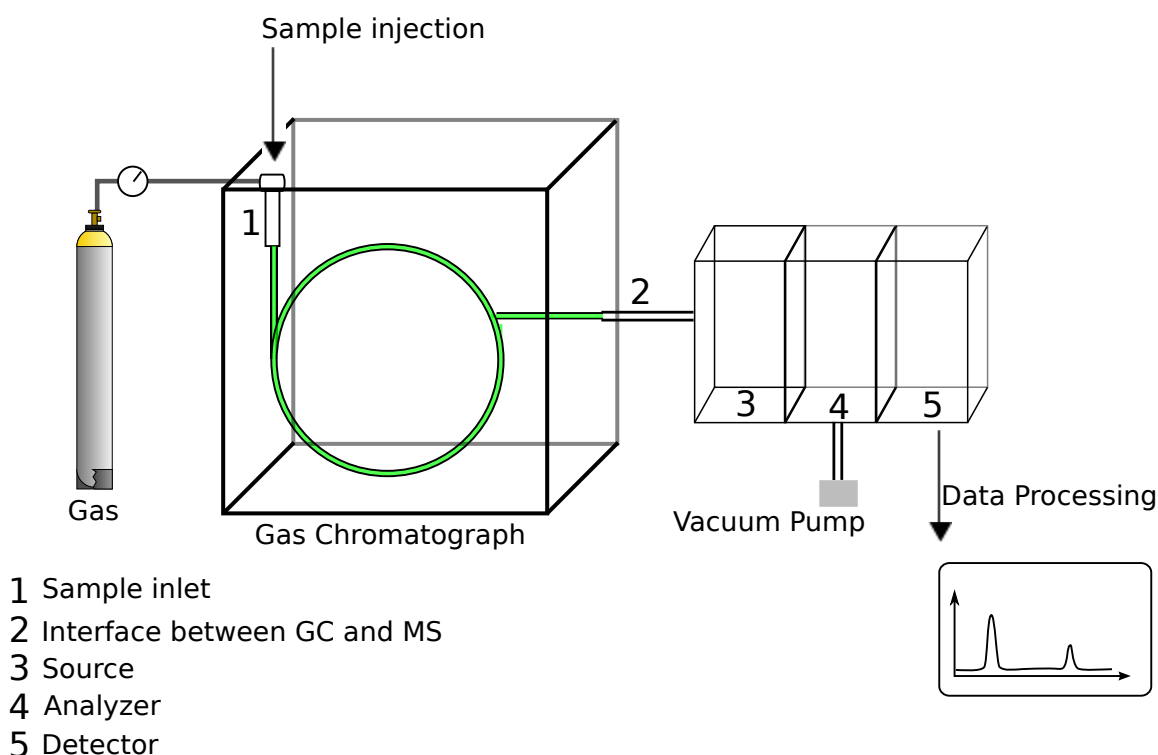


FIGURE 1.1: GC/MS scheme

For most practical purposes, the two major analytical platforms for measuring metabolite levels are MS [Lei et al., 2011] and NMR [Holmes, Nicholls, Lindon, Ramos, Spraul, Neidig, Connor, Connelly, Damment, Haselden, and Nicholson, Holmes et al.]. Coupled to a chromatographic separation technique like gas chromatography (GC) [Koek et al., 2006] or liquid chromatography (LC) [Nordström et al., 2008, Want et al., 2005], MS offers a much higher sensitivity compared to NMR. On the other hand, NMR yields specific positional information, thus complementing the information gained by MS [Schroeder et al., 2007]. Besides the classical GC and LC separation, ultra performance liquid chromatography (UPLC) [Patterson et al., 2008] and capillary electrophoresis (CE) [Lapainis et al., 2009] are widely used. Since all algorithms and methods presented in this thesis are optimized for GC/MS, I will focus on GC/MS here. Figure 1.1 depicts the rough scheme of a GC/MS instrument. Such a device is composed of two major building blocks: the GC and the MS. The GC separates compounds within a complex mixture

and the MS then subsequently ionizes and detects the mass to charge ratios (m/z) of those compounds. The most widespread ionization technique used in GC/MS is electron

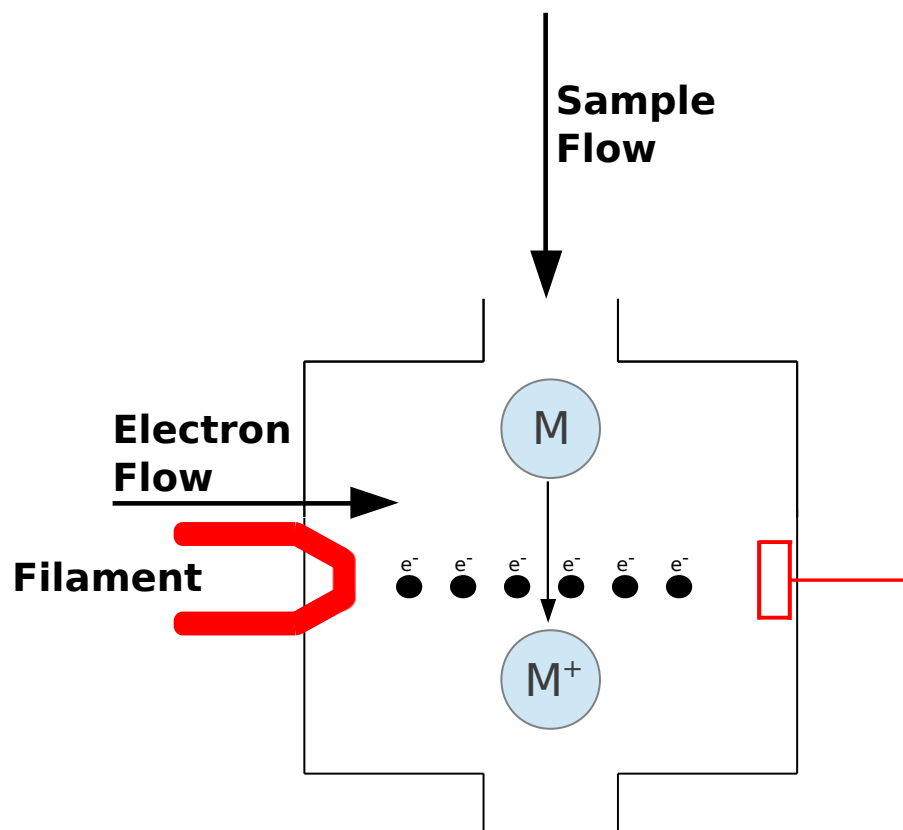


FIGURE 1.2: **Electron ionization source.** Electrons are emitted by a heated filament and accelerated towards a anode by an appropriate potential. Typically, an energy of 70 eV is used to generate a constant beam of electrons. When an electron hits a neutral sample molecule, it knocks out one of its electrons, which induces vibrations, rotations, and molecular rearrangements. As a result the molecule fragments.

ionization (EI), formerly called electron impact ionization. Figure 1.2 depicts the scheme of an ion source. Compounds are exclusively ionized in the gas phase under vacuum to form positive radical ions:



where M is a gaseous molecule, e^{-} is the electron and $M^{\bullet+}$ is the resulting radical cation of M (also called molecular ion). Since the resulting positive radical ion is highly unstable, in most cases it cannot be detected. However, EI creates reproducible fragmentation patterns, which are characteristic for a given compound. The fragmentation of gas phase ions is a complex and often hard-to-predict process. A detailed description can be found elsewhere [McLafferty and Turecek, 1994]. Although the whole fragmentation process can be very complex, there are only a few basic types of reactions that break or form chemical bonds:

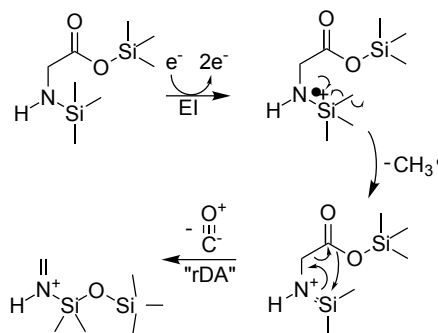


FIGURE 1.3: **Example of rearrangement after EI.** Proposed fragmentation mechanism of *N,O*-bis-(trimethylsilyl)-glycine. After expulsion of a methyl radical by alpha-cleavage next to the nitrogen, carbon monoxide loss occurs by a retro-Diels-Alder-like reaction.

1. σ -ionization: Immediately breaks a bond (affecting mostly hydrocarbons)
2. α -cleavage: A new bond is formed from a radical site and an adjacent bond is homolytically cleaved
3. Charge-induced heterolytic cleavage: Cleavage of a bond next to a charge-site
4. Rearrangements: Migrations of atoms or groups of atoms (Figure 1.3)
5. Displacement of atoms or groups of atoms
6. Eliminations

Once the compound molecules are ionized their mass to charge ratios are detected. Figure 1.4 depicts a typical mass spectrum of one GC/MS fragment ion. These fragment ions, also called isotope clusters, give rise to multiple peaks in the mass spectrum because of naturally occurring stable isotopes. While, for example, 99% of the naturally occurring carbons are ^{12}C , 1% are ^{13}C , creating these adjacent group of peaks in the mass spectrum of a fragment ion. Every peak in a fragment's mass spectrum corresponds to the same elemental composition, but different isotopic composition. The mass spectrum of a compound consists of several different isotope clusters, depending on the strength of the fragmentation during the ionization process. In conclusion, a compound's mass spectrum is mainly determined by two things:

- The fragmentation after ionization
- The elemental composition of those fragment ions

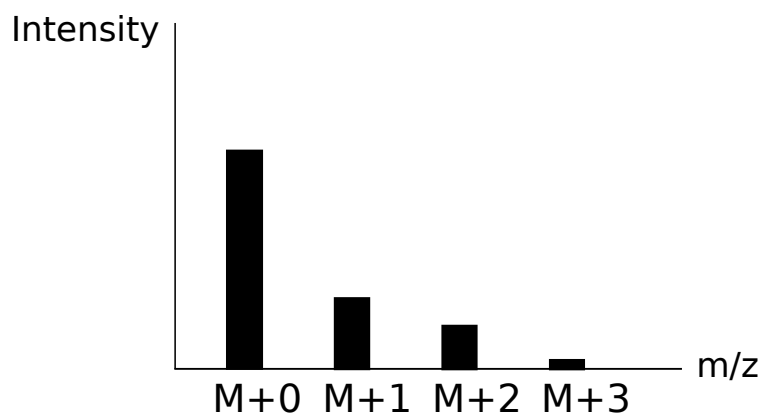


FIGURE 1.4: **GC/MS isotope cluster.** In most cases, the highest peak within an ion cluster originates from a straight combination of the lightest isotopes of all elements, also called monoisotopic peak (M+0). In a given isotope cluster, all peaks are denoted in the relationship of masses relative to the mass of the monoisotopic peak. For example, the peak with one mass unit above the monoisotopic peak is denoted M+1. In case of isotope clusters containing elements with a naturally high abundance of heavier isotopes, such as chloride or bromide, the monoisotopic peak might not be the peak with the highest intensity within the isotope cluster.

1.1.1.3 Derivatization

Most metabolites are not volatile enough to be analyzed directly by GC/MS. For that reason, polar groups such as -CO, -COOH or -NH₂ are chemically modified prior to analysis. There exist a wide range of derivatization agents, however silyl derivatives are suited best for GC/MS analysis. In this thesis I used mainly *N*-Methyl-*N*-(trimethylsilyl)-trifluoroacetamide (MSTFA) which creates the typical trimethylsilyl (TMS) derivatives and *N*-(*tert*-butyldimethylsilyl)-*N*-methyltrifluoroacetamide (MTBSTFA) which creates the typical *tert*-butyldimethylsilyl (TBDMS) derivatives (Figure 1.5).

1.1.2 Methods Of Compound Identification

One of the advantages of GC/MS and electron ionization is relatively easy compound identification (compared to e.g. LC/MS), because electron ionization generates reproducible and characteristic mass spectra. These mass spectra can be collected and stored in a reference library, which can then be used to identify compounds of GC/MS measurements. To do this, a spectrum similarity score between the mass spectrum of a measured compound (S_{mes}) and *all* mass spectra in the reference library (S_{lib}) is calculated. The library compound with the highest spectrum similarity score is then assigned to the measured compound. Additionally, a score based on the retention time can be calculated to discriminate compounds that have highly similar mass spectra. Usually, both

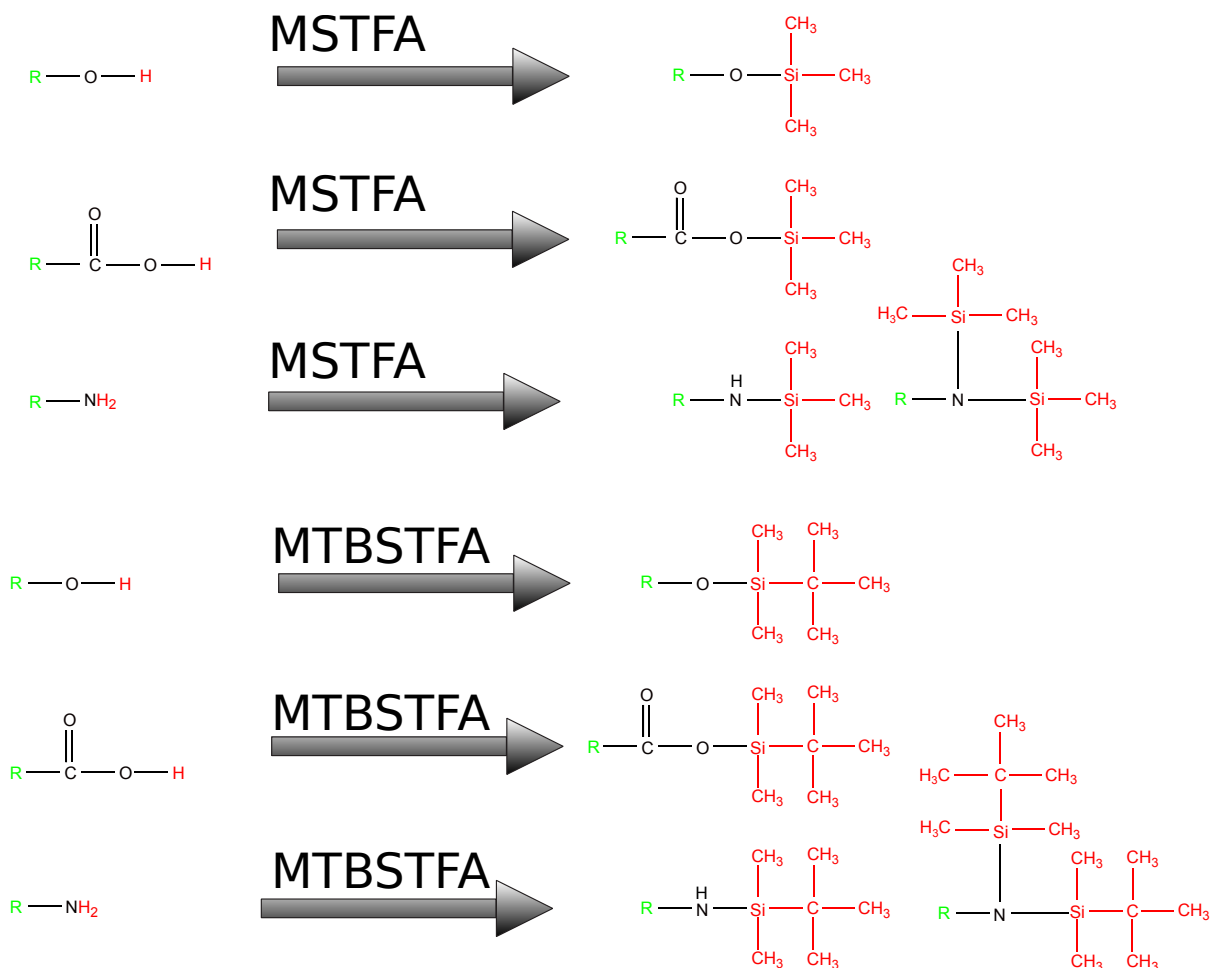


FIGURE 1.5: **Schematic of GC/MS derivatization.** Hydrogens of polar groups that are substituted either by a trimethylsilyl (TMS) group or a tert-butyldimethylsilyl (TBDMS) group are shown in red. A wide range of organic compounds can be easily derivatized to their respective TMS and TBDMS derivatives, which has the advantage that silyl derivatives are generally less polar, more volatile and thermally more stable than their precursors.

scores are applied to align compounds of a batch of chromatograms in a metabolomics experiment.

1.1.2.1 Spectrum Similarity Score

In the past decades several spectrum similarity based identification algorithms have been developed. Of these the weighted dot product has proven to perform best in terms of accuracy [Stein and Scott, 1994]. A compound's mass spectrum S is defined as a set of pairs of masses and intensities:

$$S = (m_1, i_1) = p_1, (m_2, i_2) = p_2, \dots, (m_n, i_n) = p_n \quad (1.2)$$

$$m_i < m_{i+1} \quad i \in N$$

According to Stein and Scott, each mass intensity pair p_i of both the measured and reference spectrum is weighted according to the following rule:

$$p^w = [mass]^a \times [intensity]^b \quad (1.3)$$

where a and b are the weighting factors that represent the contribution of the m/z value and the peak intensity, respectively. Stein and Scott reported that the optimal values are $b = 0.6$ for intensity scaling and $a = 3$ for mass weighting [Stein and Scott, 1994]. This way, the relative influence of minor intensities at higher masses is increased. However, several different values have been proposed to be the optimal weighting factors over the last years. Recently, Kim *et al.* showed that weighting factors should be chosen individually based on the reference library used [Kim et al., 2012].

We denote a set of weighting factors as $w=(a,b)$. The dot product of the library (S_{lib}^w) and measured (S_{mes}^w) spectrum is then calculated as follows:

$$Score_{SpecDot} = \frac{S_{lib}^w \times S_{mes}^w}{\|S_{lib}^w\| \|S_{mes}^w\|} \quad (1.4)$$

As a result, the spectrum similarity score is located within the interval zero (no identity) to one (identical spectra).

1.1.2.2 Retention Index Similarity Score

Because the retention time is dependent on the instrument used, the GC-capillary, or the applied temperature program, etc., we use the Kovats [Kovats, 1958] retention index (RI) for all retention time-based similarity measures. Assuming that the determined retention indices for a certain compound are distributed in a Gaussian manner across different chromatograms, a Gaussian function is used for the RI based similarity index calculation:

$$Score_{RI} = e^{-\frac{(ri_{lib}-ri_{mes})^2}{2x^2}} \quad (1.5)$$

1.1.2.3 Chromatogram Alignment

If a number of GC/MS chromatograms are to be analyzed comparatively, it is necessary to align similar compounds among the different chromatograms. To additionally account for the retention time of a compound, a combined similarity score $Score_{total}$ based on the spectrum and retention index similarity is calculated. Because the spectral profile of a compound contains more information than the retention index, it is weighted stronger

$$Score_{total} = \sqrt[3]{Score_{spec}^2 \cdot Score_{RI}} \quad (1.6)$$

1.1.3 Targeted vs Non-Targeted Metabolomics

Although metabolism was extensively studied over the last decade and many pathways and their respective metabolites have been uncovered, the assumption that this knowledge is complete is probably false, especially for disease specific alterations of metabolism. However, a targeted approach relies partially on this assumption, because it only takes a predefined set of metabolites into consideration. Apparently, this set does not include unknown metabolites but can include metabolites not known to be produced by a specific organism or under a specific condition. A non-targeted approach, however, tries to analyze all measurable metabolites. As such, a non-targeted approach is able to capture metabolites not yet known or not known to be produced by a specific organism or under a specific condition. The advantage of a targeted approach is that it can obtain absolute quantitative information. For that reason a non-targeted approach cannot replace a targeted approach, but can yield additional information that is disguised by a targeted approach. To obtain a more comprehensive view of cellular metabolism, both techniques should be combined, with the targeted approach being preceded by the non targeted approach. This way the non targeted approach works as a discovery tool to better constrain the targeted analysis.

1.2 Stable Isotope Assisted Metabolomics

The above described technologies provide a tool set to conduct metabolomics research. Although they can detect changes in metabolite concentrations, no information about increases or decreases of fluxes in the associated pathways can be obtained. However, metabolic fluxes reflect the quantitative endpoint of the interplay between gene expression, protein synthesis, post-translational modifications and thermodynamic constraints, therefore representing the cell's final phenotype. Analyzing intracellular metabolic fluxes is essential investigating the physiological state of a cell and thereby revealing disease specific alterations or dysregulations of metabolic conversion rates and enzyme activities. As a consequence of the complex regulation of metabolic pathways, significant flux changes are sometimes associated only with a modest change in metabolite concentrations [Fell, 2005]. To obtain information about intracellular dynamics, stable isotope labeled components can be applied. For that, stable isotope (e.g. ^{13}C , ^{15}N) labeled substrates (e.g. glucose, glutamine) are fed to the target system (e.g. cell culture, tissue, whole organism) until cellular metabolism distributes the isotopes throughout the metabolic network. Based on the reaction rates and enzyme activities present in the system, distinct labeling patterns will arise affecting the molecular weight of the fragment ions.

1.2.1 Mass Isotopomer Distribution

Using an MS approach, fragment ions are separated according to their m/z ratio. Based on the incorporated labeled atoms, the m/z ratio is shifted by one or more atomic mass units, resulting in different shifted mass spectra. According to the international union of pure and applied chemistry (IUPAC), an isotopomer is defined as an isomer having the same number of each isotopic atom but differing in their positions. On the other hand, an isotopologue is defined as a molecular entity that differs only in isotopic composition (number of isotopic substitutions). Within the metabolic community the term mass isotopomer is used as a synonym for isotopologue. The number of isotopomers (2^n) is therefore always larger than the number of mass isotopomers or isotopologues ($n+1$), where n is the number of possible isotopic substitutions (Figure 1.6). Based on the MS measurement, mass isotopomer distributions (MIDs) as the relative amount of each mass isotopomer can be calculated by solving a linear equation system [Lee et al., 1991]. In order to set-up the equation system and to calculate correct MIDs, a correction matrix is needed that corrects the MIDs for naturally occurring stable isotopes. Figure 1.6b depicts the correction matrix for a two carbon compound. The first column corresponds to the natural mass isotopomer distribution of the unlabeled compound (M_{00} , M_{01} and M_{02} in Figure 1.6). The second column corresponds to the natural mass isotopomer distribution if one of the ^{12}C is replaced with a ^{13}C (M_{10} , M_{11} and M_{12} in Figure 1.6). The third column corresponds to the natural mass isotopomer distribution if both of the ^{12}C are replaced with a ^{13}C (M_{20} , M_{21} and M_{22} in Figure 1.6). This correction matrix can be setup by predicting the mass spectrum of the tracer using multinomial expansion based on the natural abundance of stable isotopes along with the chemical formula of the fragment ion. If the chemical formula of the fragment ion is not known, the unlabeled reference spectrum of the fragment ion can be applied [Jennings and Matthews, 2005]. In addition to static metabolite concentrations, MIDs provide dynamic information about how a compound is metabolized within the cell. This includes the involved pathway(s), enzyme activities and the fate of the labeled atoms (assuming the respective pathway intermediates can be measured). The importance of measuring MIDs is best explained by a short example: Glucose can be metabolized to pyruvate either via glycolysis or the pentose phosphate pathway (PPP). Depending on the respective pathway the carbon-carbon bonds of glucose are broken and rearranged specifically. If glucose is metabolized via the PPP, the first carbon atom is decarboxylated in the oxidative phase, whereas it is conserved in glycolysis (Figure 1.7). As a consequence, the relative amount of glucose metabolized through the PPP compared to glycolysis can be revealed by using a glucose tracer labeled on the first and second carbon atoms ($1,2\text{-}^{13}\text{C}_2$ glucose). The MID of pyruvate gives direct information of the metabolic flux of glucose through the two pathways. Pyruvate containing one stable carbon isotope (M_1) was metabolized via

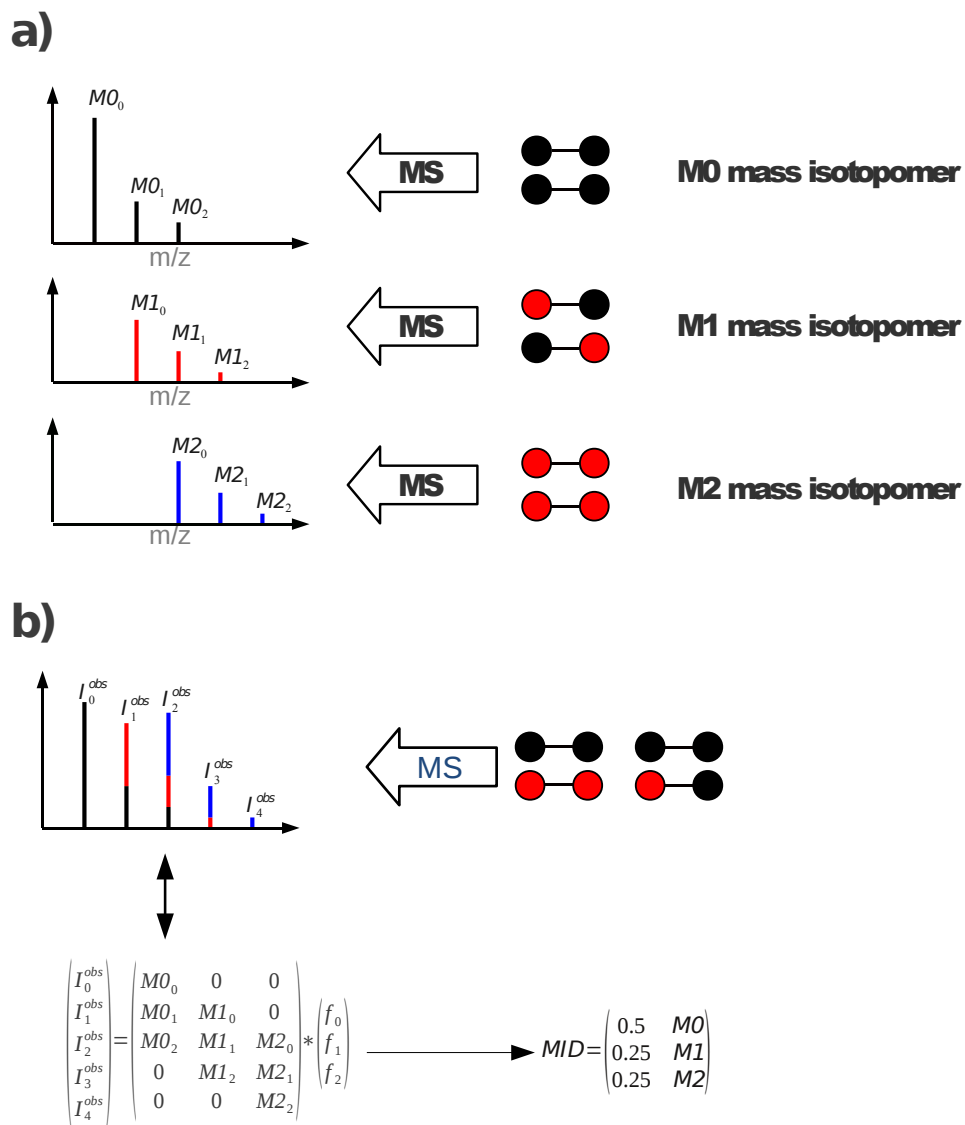


FIGURE 1.6: Mass isotopomer distribution (MID). (a) Based on the number and position of incorporated labeled atoms, different isotopomers are shown (^{12}C in black, ^{13}C in red). The mass spectrum is depicted along with the corresponding mass isotopomer. (b) In reality only one mass spectrum is measured for a complex mixture of mass isotopomers and the fraction of each mass isotopomer has to be determined by solving the linear equation system. In this case 50% of the molecules are unlabeled, 25% contain one stable isotope and 25% contain two stable isotopes (Figure based on [Hiller et al., 2011])

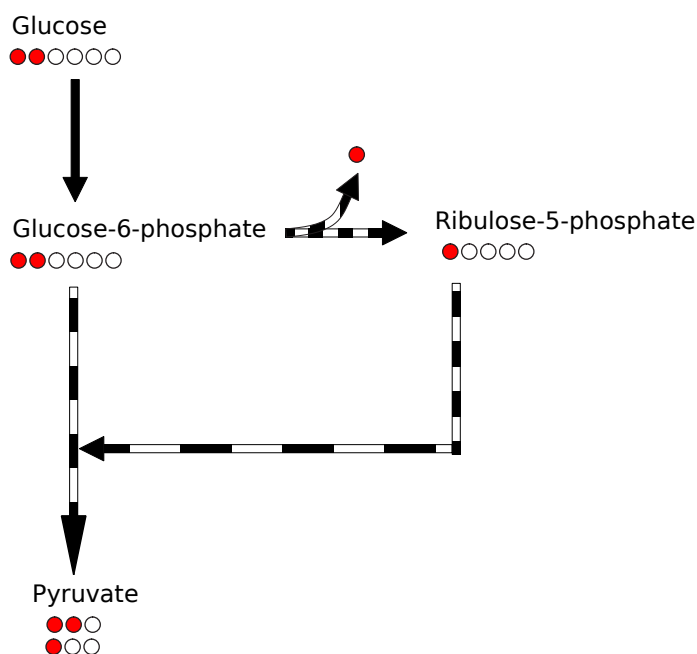


FIGURE 1.7: **Simplified scheme of the pentose phosphate pathway and glycolysis.** The carbon backbones of every metabolite is shown in circles (^{12}C in white and ^{13}C in red). Since the first carbon atom originating from glucose is decarboxylated in the oxidative phase of the pentose phosphate pathway (PPP), whereas it is conserved in glycolysis, the number of ^{13}C atoms in Pyruvate yield information about the activity of the two pathways. Pyruvate containing one ^{13}C atom was metabolized via PPP and pyruvate containing two ^{13}C atoms via glycolysis.

PPP and pyruvate labeled by two stable isotopes (M2) via glycolysis. The ratio of M1 and M2 mass isotopomers represent the respective flux through PPP and glycolysis.

1.3 Metabolic Flux Analysis

MFA aims to quantify all intracellular fluxes in a given system. Initial methods were based solely on a known stoichiometry of the biochemical reaction network of interest [Stephanopoulos, 1999, Varma and Palsson, 1994]. In that context, intracellular fluxes can be inferred by measuring the metabolic input and output under the assumption of a metabolic steady-state. An example for a very simple reaction network is depicted in Figure 1.8a. In total this reaction network has 5 fluxes. The steady-state constraint yields the following flux relations:

$$u = v + w \quad v = x \quad w = y \quad (1.7)$$

If the extracellular fluxes u and x are measured, then the fluxes v , w and y can be calculated using the following equations:

$$v = x \quad w = u - x \quad y = u - x \quad (1.8)$$

However, in living cells intracellular metabolism is much more complex than illustrated in the example. Particularly, stoichiometric MFA fails in the following situations: bidirectional reaction steps and certain parallel or cycle reactions [Wiechert, 2001a]. Figure 1.8b depicts the same reaction network, but with one additional reaction r . For this reaction network intracellular fluxes cannot be inferred from measuring the extracellular fluxes alone. To overcome these limitations more sophisticated methods such as ^{13}C -MFA evolved.

Figure 1.9 depicts an overview of the experimental and computational steps necessary to perform ^{13}C -MFA. The first important step is to select a suitable isotopic tracer. The precursor and the position of the label should be chosen carefully, because this heavily influences the precision and accuracy of the flux estimation. For the example illustrated in Figure 1.8b, only a label on the third or fourth carbon atom yields additional information in which case the flux r can be determined by the percentage of the labeled carbons in metabolite B. A recent study evaluated different ^{13}C tracers for their use in ^{13}C -MFA experiments [Metallo et al., 2009]. For example, the 1,2- $^{13}\text{C}_2$ -Glucose tracer is suited best to study the PPP and glycolysis, whereas 1- ^{13}C -glutamine is suited best to study the reductive flux from α -ketoglutarate to citrate. The second important step is to create a metabolic network model that defines the metabolic reactions and atom transitions for the pathway of interest. Once the isotopic tracer and the metabolic network model are defined, a stable isotope labeling experiment is performed. For that, ^{13}C -labeled precursors are introduced into the network. The redistribution of the label into other metabolites is measured after the system reaches an isotopic and metabolic steady-state assuming constant intracellular fluxes and labeling patterns. The labeling patterns or MIDs can then be detected either by mass spectrometry or NMR. Since intracellular fluxes cannot be measured directly, ^{13}C -MFA estimates the cell's flux state based on measured MIDs. For that, an iterative non linear least squares fitting procedure is applied to find a set of fluxes that account best for the observed MIDs [Wiechert, 2001b]. In order to do that, a mathematical model is required that can simulate MIDs for a given set of steady state fluxes.

During the last years, several different approaches have been developed to mathematically describe the relation between MIDs and the corresponding fluxes. Schmidt *et al.* used a matrix approach [Schmidt et al., 1997] to generate a complete set of isotopomer balances, based on the idea of atom-mapping matrices (AMM) [Zupke and Stephanopoulos, 1994]. Generalizing the concept of AMMs, which track the transfer of carbon atoms

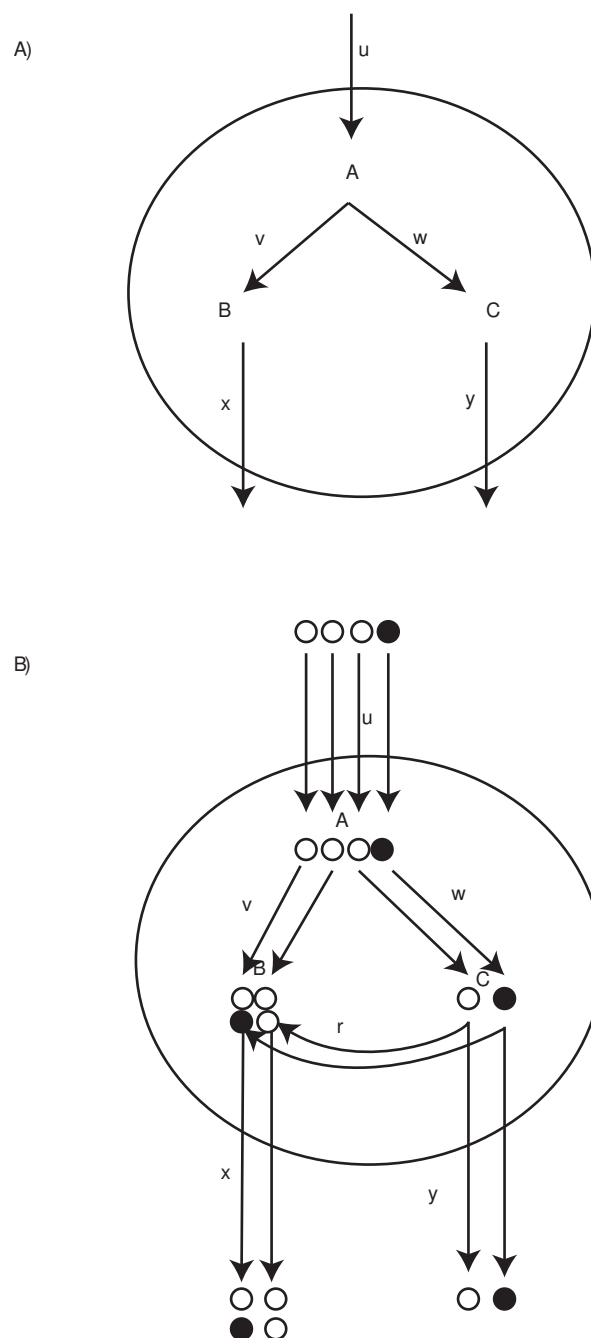


FIGURE 1.8: (a) A simple example network, where intracellular fluxes can be calculated with stoichiometric MFA as shown in the text. (b) The same reaction with one additional intracellular reaction r . The carbon transitions are shown for every compound (^{12}C in white and ^{13}C in black). Based on the carbon transition network the flow of label from compound A can be traced to the metabolites B and C to infer the flux r .

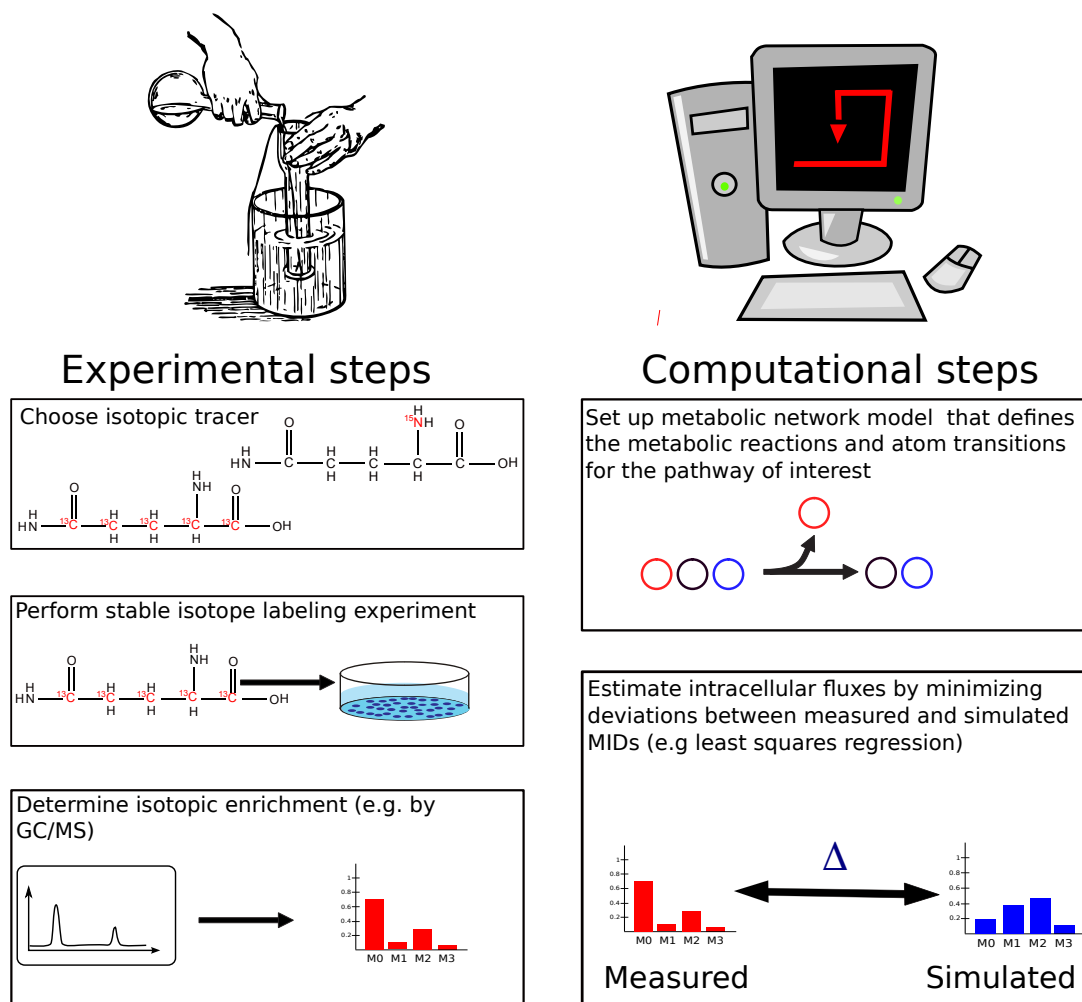
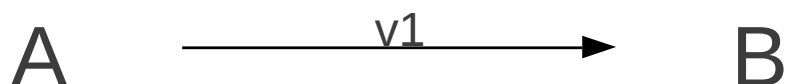


FIGURE 1.9: MFA overview

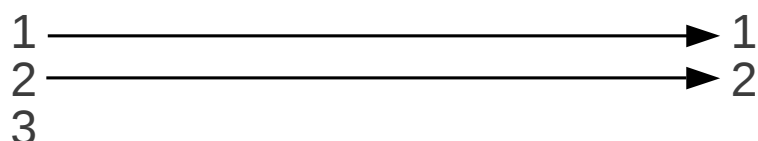
from reactants to products, Schmidt introduced isotopomer mapping matrices (IMMs). In this case, isotopomer distribution vectors (IDVs) are generated for each metabolite, which contain the molar fraction of each isotopomer. IMMs are then used to sum up the fraction of each reactant isotopomer that form a respective product isotopomer (Figure 1.10). Later, Wiechert *et al.* further extended this idea and introduced the concept of cumulative isotopomers (cumomers) [Wiechert *et al.*, 1999]. Cumomers balances can be calculated computationally more efficient than isotopomer balances [Wiechert *et al.*, 1999], but they do not reduce the size of the problem meaning that there are always as many cumomer balances as there are isotopomer balances. To overcome this limitation, Antoniewicz *et al.* developed the elementary metabolite unit (EMU) framework [Antoniewicz *et al.*, 2007]. Here, the minimal amount of information is calculated fully describing the measured labeling states. The authors claim that the number of variables are an order of magnitude smaller compared to the previously described techniques. It is important to note that the size of the defined metabolic network is a critical factor

a)

Reaction stoichiometry

**b)**

Atom transitions

**c)**









Atom mapping matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} * \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$$

d)

Isotopomer mapping matrix

Isotopomers binary state

	000
	001
	010
	100
	011
	101
	110
	111

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} * \begin{pmatrix} A_{000} \\ A_{001} \\ A_{010} \\ A_{100} \\ A_{011} \\ A_{101} \\ A_{110} \\ A_{111} \end{pmatrix} = \begin{pmatrix} B_{00} \\ B_{01} \\ B_{10} \\ B_{11} \end{pmatrix}$$





	00
	01
	10
	11

FIGURE 1.10: (a) A simple reaction network model used as an example to create an atom mapping matrix (AMM) and an isotopomer matrix (IMM). The network contains one reaction $v1$. (b) Atom transition network, describing the carbon atom transitions from compound A to B occurring in reaction $v1$. (c) Atom mapping matrix and metabolite vector that describes the reaction network and atom transitions. (d) Isotopomer mapping matrix that converts the isotopomer distribution vector of compound A to the isotopomer distribution vector of compound B. The possible isotopomers are shown in circles (^{12}C in white and ^{13}C in black) with the corresponding binary code.

in ^{13}C -MFA. As the network increases in size, more constraints (measured MIDs) are needed to perform ^{13}C -MFA. On the other hand, a small network may be not able to capture the complex metabolic network present in living cells. In Chapter 3, I will introduce a novel algorithm that helps to incorporate more constraints and with it a bigger metabolic network to perform ^{13}C -MFA.

Initially, MIDs were derived by only measuring proteogenic amino acids [Zamboni et al., 2009]. However, the time for biomass to reach an isotopic steady-state is at least one cell generation time [Wiechert and Nöh, 2005]. Hence, a long and cost-intensive experimental duration is necessary. A more direct and intuitive way is to directly measure MIDs for intracellular pathway intermediates of interest, though, this is challenging for two reasons: First, the concentrations of intermediates in central carbon metabolism are usually very low. Second, the metabolic turnover rates can be very high. However, recent technological improvements in sample preparation and mass spectrometry have partly overcome these limitations and allow accurate determination of MIDs for many intracellular metabolites [Antoniewicz et al., 2007].

^{13}C -MFA has been applied to calculate fluxes in various systems, such as bacteria, yeast, plants and mammalian cells [Marin et al., 2004, Masakapalli et al., 2010, Niklas and Heinze, 2012, Niklas et al., 2010, Raghevedran et al., 2004].

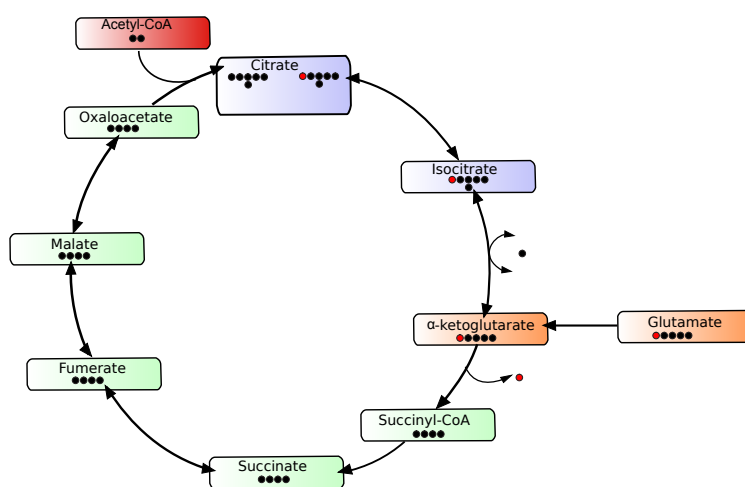


FIGURE 1.11: **Simplified TCA cycle to demonstrate the value of stable isotopes.** The carbon backbones are shown in red for ^{13}C and in black for ^{12}C . By following the isotopic labeled atoms of glutamate the reversibility of the reactions converting α -ketoglutarate to citrate becomes clear. Moreover, the reductive flux from α -ketoglutarate to citrate can be calculated.

In summary, stable isotopes can add value to metabolomics studies in two ways: (i) Stable isotopes can reveal unanticipated reactions that are currently not known or not associated for a given metabolic state. (ii) Stable isotopes can be used to calculate

absolute values for intracellular fluxes. The network in Figure 1.11 illustrates a simplified version of the TCA cycle. The reductive and oxidative flux from α -ketoglutarate to citrate cannot be distinguished by only measuring the metabolite concentrations of e.g. citrate, but can be inferred from the MIDs.

1.4 Aim And Outline Of Thesis

In the emergent field of systems biology metabolomics has become a key player. Specifically, non-targeted metabolomics methodologies have proven to be indispensable. Recent technological improvements in sample preparation and mass spectrometry have provided the means to detect an increasing number of intracellular metabolites. However, the ability to extract biological knowledge out of these data mainly relies on the applied computational analysis. For that reason, novel algorithms have to be developed to obtain more biological knowledge out of the flood of metabolomics data. Moreover, the appropriate software tools have to be developed to facilitate the use of these algorithms within the metabolomics community.

- **Chapter 2** describes the development and application of non-targeted metabolomics methodologies. First, an alternative spectrum similarity measure that is based on the specific fragmentation patterns of electron impact mass spectra is presented. I developed the isotope cluster based compound matching (ICBM) to overcome the problem of mismatched compounds typically occurring during chromatogram alignment. The ICBM algorithm allows a sensitive peak detection step without losing the specificity of the compound matching. As such, the algorithm is most efficient for the alignment of compounds across different chromatograms. Specifically for non-targeted analyses, the ICBM algorithm outperforms conventional tools as for example the dot product. Moreover, this chapter includes the application of the ICBM algorithm to characterize the metabolome of the human mesencephalic cell line (LUHMES, Lund human mesencephalic) [Scholz et al., 2011] under different oxygen conditions.
- **Chapter 3** describes a methodology for the determination of chemical formulas and retained atoms for mass spectral fragment ions. Chemical formulas usually form the basis of MID calculations for specific ions. Hence, the correct assignment of chemical formulas to fragment ions is of crucial importance for the calculation of accurate MIDs. Furthermore, the retained carbon atoms of fragment ions are necessary to perform ^{13}C -MFA. However, the process of mass spectral fragmentation is complex and assigning chemical formulas and retained atoms to mass

spectral ions is non-trivial. To address this shortcoming, I developed an approach, based on a systematic bond cleavage, to determine chemical formulas and the retained atoms for GC/MS fragment ions. I applied the fragment formula calculator (FFC) to determine the chemical formulas for a wide range of TMS and TBDMS derivatized fragment ions.

Chapter 2

Non-Targeted Metabolomics Methodologies

This chapter covers my contributions to publications about non-targeted metabolomics methodologies. This includes the spectrum similarity measure “Isotope Cluster-Based Compound Matching” and a new software tool for the non-targeted detection of stable isotope labeled compounds. As an example to point out the importance of non-targeted metabolomics methodologies, I will discuss the discovery of the previously unknown link between immunoresponsive gene 1 (*Irg1*) and itaconic acid. The main results are summarized with respective cross references to already published articles. For methodological details, please consult the corresponding manuscript [[Hiller et al., 2013](#), [Michelucci et al., 2013](#), [Wegner et al., 2013](#)].

Wegner, A.; Sapcariu, S. C.; Weindl, D.; Hiller, K.

Analytical chemistry (2013), 85(8), 4030-4037

Hiller, K.; **Wegner, A.**, Weindl, D.; Cordes, T.;

Metallo, C. M.; Kelleher, J. K.; Stephanopoulos, G.

Bioinformatics (2013), 29(9), 1226-8

Michelucci, A.; Cordes, T.; Ghelfi, J.; Pailot, A.;

Reiling, Npublication given in appendix.; Goldmann, O.; Binz, T.; **Wegner, A.**;

Tallam, A.; Rausell, A.; Buttini, M.; Linster, C.;

Medina, E.; Balling, R.; Hiller, K.

PNAS (2013), 110(19), 7820-7825

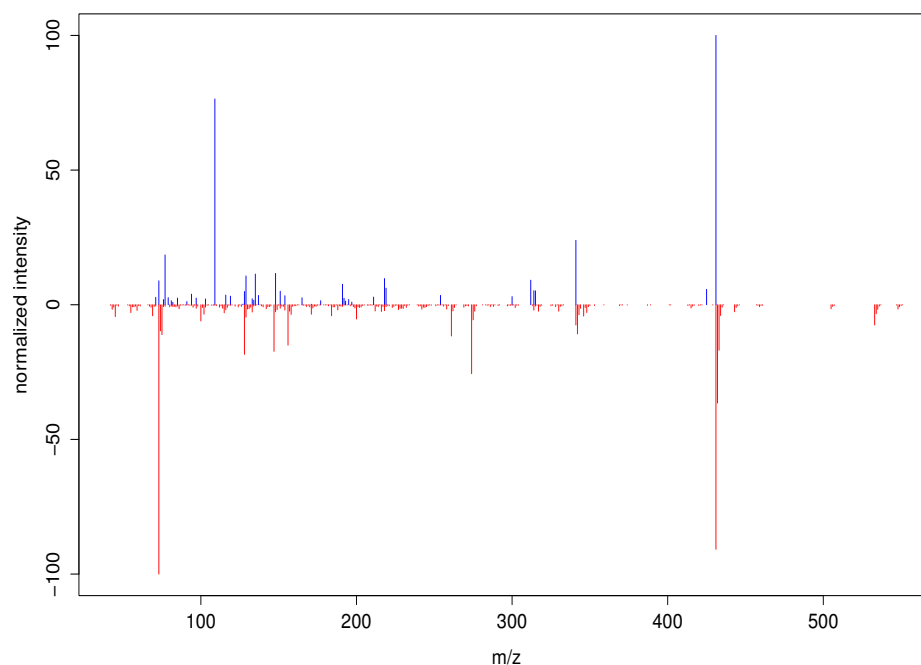


FIGURE 2.1: **Example of a spectrum similarity score calculated with the dot product.** The measured spectrum is shown in blue in the positive direction and the matched spectrum of the library is shown in red in the negative direction. The library spectrum obviously does not match the library spectrum. However, the spectrum similarity score calculated with the dot product is 0.9, which is relatively high. One reason for the high score is the fact that single peaks at higher masses have a bigger influence on the final spectrum similarity score compared to peaks at lower masses. The peak at mass 420 in the measured spectrum dominates the final spectrum similarity score, which leads to a false positive identification.

2.1 Isotope Cluster Based Compound Matching

A typical comparative metabolomics analysis (both targeted and non-targeted) consists of three steps: First, compounds (clusters of ion-chromatographic peaks) are detected in every measured chromatogram. Second, detected compounds are matched across all chromatograms (chromatogram alignment) and quantitative values are calculated. Third, matched quantitative values are statistically analyzed (e.g. principal component analysis, self-organizing maps, etc.). While in the beginning metabolomics studies mainly focused on the quantification of a targeted set of previously known metabolites, recent studies have tried to quantify all detectable metabolite peaks within a chromatogram. This non-targeted approach, however, generates a bottleneck already at the first step of analysis. Metabolites of low concentration will be hard to distinguish from “noise peaks” (usually small peaks near the GC baseline) in the compound detection step. The more sensitive the peak detection step, the more erroneously detected compounds are present in the data set, making it more difficult to match “real” chromatographic peaks across different samples. On the other hand, “real” chromatographic

peaks might be overlooked when less sensitive settings are applied. Finding the right tradeoff between sensitivity and specificity for the compound detection can be challenging. For a targeted approach, compounds of interest are known in advance, and optimal compound-specific settings can be determined by evaluating the results for these compounds. However, in a non-targeted methodology, such evaluation is not possible, and compounds of interest might be inadvertently removed when less sensitive compound detection settings are applied. Therefore, settings with a high sensitivity should be applied in these cases.

Current spectrum-matching-based identification algorithms, such as the dot-product, struggle with data generated by non-targeted metabolomics experiments, when a highly sensitive compound detection step was applied. One major problem of the dot-product is that, due to the applied scaling algorithm, single peaks at higher masses can falsify the spectrum similarity score which leads to wrongly assigned compounds in the chromatogram alignment step (see Figure 2.1). Consequently, quantitative values for these mismatched compounds are calculated incorrectly which may conceal an important result or dissembles a wrong result of this experiment. To overcome this limitation I have developed the isotope cluster based compound matching (ICBM) algorithm [Wegner et al., 2013].

The ICBM algorithm places a higher emphasis on the specific fragmentation pattern of EI mass spectra and takes into account the natural stable isotope abundances. The fragmentation pattern and the distribution of natural stable isotopes are the two most characteristic features of EI mass spectra and are, therefore, well suited to discriminate between different mass spectra. The isotope cluster based matching algorithm consists of four main steps (Figure 2.2):

1. All mass spectral isotope clusters are determined for the measured and the reference spectrum respectively.
2. Isotope clusters are aligned based on the m/z values of their monoisotopic peaks.
3. Two similarity scores are calculated. For isotope clusters with matching monoisotopic peaks, a score based on the isotope cluster's peak ratios is calculated. For non-matching isotope clusters and non-grouped peaks, a score based on the dot product is calculated.
4. The two similarity scores are combined to a final score between 0 and 1, where 1 represents a perfect match and 0 a mismatch between the query and the reference spectrum.

In contrast to the dot product, the ICBM algorithm focuses more on those features of a compound's mass spectrum that are characteristic for the compound. In this light, the

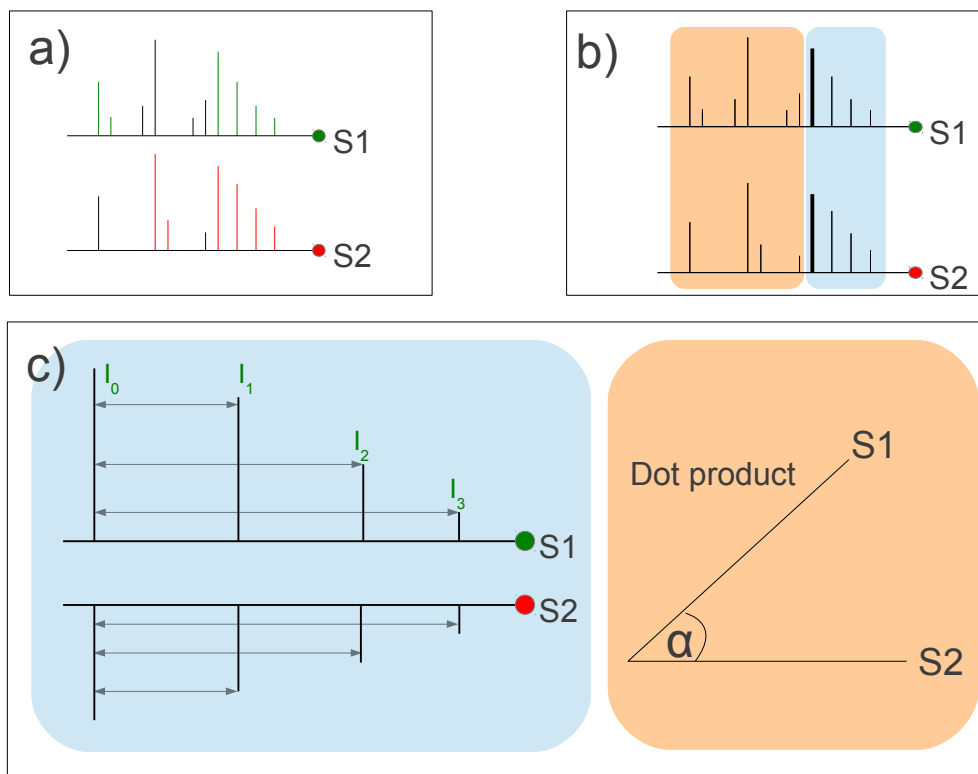


FIGURE 2.2: **Overview of the ICBM algorithm.** (A) Detected isotope clusters of the measured spectrum are colored in green and of the library in red. (B) Detected isotope clusters are aligned based on the masses of their monoisotopic peaks. Two isotope clusters are considered a match if the mass of their monoisotopic peak is identical (shown in bold). After the alignment the peaks of both spectra are divided in two separate subsets. Peaks of identical isotope clusters are shown in blue, and peaks of non-matching isotope clusters or not grouped within a fragment in orange. (C) Two similarity scores based on the fragment alignment are calculated and combined.

ICBM algorithm works in part analogous to a manual inspection of the mass spectrum by an expert.

2.1.1 Mathematical Description Of The ICBM Algorithm

I define an ion cluster as a subset of a spectrum S :

$$f = \{p_k, \dots, p_l\}, 0 < k < l \leq n, f \subset S \quad (2.1)$$

where p_k denotes the first peak and p_l the last peak of the ion cluster. A spectrum S can have multiple ion clusters, which are all disjointed subsets of S . As a reference point within an ion cluster, I use the peak with the highest intensity (I_M). I will use the term isotope cluster normalization to refer to the ratio r_i of each isotope cluster's

peak intensity in relation to the intensity of its peak I_M :

$$r_i = \frac{I_{M+i}}{I_M} \quad k \leq i \leq l \quad (2.2)$$

2.1.1.1 Isotope Cluster Determination

The algorithm for the isotope cluster detection iterates through S once. All consecutive peaks with a mass difference of one unit and decreasing intensities are grouped together into separate isotope clusters:

$$m_j - m_{j+1} = -1 \wedge I_{M_j} > I_{M_{j+1}} \quad (2.3)$$

In case of overlapping isotope clusters or an isotope cluster containing elements with a high abundance of natural stable isotopes such as chlorine or bromine, the algorithm splits them into two separate isotope clusters.

2.1.1.2 Isotope Cluster Alignment

Isotope clusters of the measured spectrum are aligned to the isotope cluster of the library spectrum based on the mass of their monoisotopic peaks (Figure 2.2b). Two isotope clusters are considered a match if the masses of their monoisotopic peaks are identical. In case of a peak at mass m_i present in one isotope cluster but not in its counterpart, a peak of mass m_i and intensity 0 is added to the corresponding isotope cluster. This way, aligned isotope clusters always have the same number of peaks. The measured and the library spectra can then be divided into two subsets of peaks. One set contains all peaks from the matching isotope clusters (F), and the other contains the remaining peaks of the spectrum (R).

$$F \cup R = S \quad (2.4)$$

In the illustrated example, peaks of the set F are highlighted in blue and peaks of the set R in orange (Figure 2.2b).

2.1.1.3 Similarity Score Calculation

On the basis of the alignment, one score is calculated for set F (matched isotope clusters) and one score for R (non-matching isotope clusters and peaks not grouped within an isotope cluster) (Figure 2.2c). First, for each matched isotope cluster in F, all peak intensities are normalized by the respective monoisotopic peak. Second, the distance

d between two isotope clusters is calculated by summing the absolute values of the differences between corresponding normalized peak intensities:

$$d = \sum_{i=0}^n |r_{lib_i} - r_{mes_i}| \quad (2.5)$$

To keep the score within the interval $[0,1]$, the contribution of one isotope cluster pair to the total similarity score is weighted by the number of isotope cluster peaks and an intensity scale. Therefore, the isotope cluster's summed intensity is divided by the total intensity of the mass spectrum to obtain the intensity fraction of the isotope cluster in S :

$$\begin{aligned} x &= \text{number of peaks within } f \\ \text{intensity scale} &= \frac{I_{fmes}}{I_{Smes}} + \frac{I_{flib}}{I_{Slib}} \\ \text{scale} &= \text{intensity scale} \cdot x \end{aligned} \quad (2.6)$$

The total isotope cluster-based distance of the two mass spectra is then calculated as follows:

$$Score_F = \frac{1}{n} \cdot \sum_{i=0}^n d_i \cdot \text{scale}_i \quad (2.7)$$

where n is the number of matched isotope clusters. This calculation transforms the distance in the interval between zero (identical spectra) to one (no identity). To make this score comparable to the score of the dot product, $Score_F$ is inverted within the interval $[0,1]$:

$$\text{Score}_F = 1 - Score_F \quad (2.8)$$

For the remaining peak set R , a similarity score based on the dot product (see equation 1.4) is calculated. These two similarity measures are then combined to form a composite spectrum similarity score. To reduce the bias to favor one of the two scores, a weighting factor w_F based on the summed intensities of all matched isotope clusters is calculated:

$$w_F = \left(\frac{I_{Fmes}}{I_{Smes}} + \frac{I_{Flib}}{I_{Slib}} \right) \cdot \frac{1}{2} \quad (2.9)$$

When for example the summed intensity of all matched isotope clusters encompasses 70% of the total intensity of the measured and the library spectrum, the isotope cluster-based score is weighted with 0.7 and the dot product score with 0.3. The final spectrum similarity score is then calculated as follows:

$$Score_{SpecIC} = (1 - w_F) \cdot Score_{SpecDot} \cdot w_F \cdot Score_F \quad (2.10)$$

2.1.2 Applications Of The ICBM Algorithm

The ICBM algorithm is most efficient for a non-targeted chromatogram alignment, which has the advantage that metabolites not present in a reference library will pop up in the analysis result as unidentified metabolites. As such, a non-targeted chromatogram alignment is able to capture metabolites not yet known or not known to be produced by a specific organism or under a specific condition. In case an unidentified metabolite pops up in the analysis, there exist several strategies that can be used to identify these compounds. One could use a different measurement technique, such as LC/MS, to infer a chemical formula or NMR to gain structural knowledge. A more straightforward way is to use a bigger reference library for the identification step. This assumes, however, that the unidentified metabolite was annotated previously. The most comprehensive EI mass spectral reference library is maintained by the National Institute of Standards and Technology (NIST) and comprises 243,893 reference spectra of 212,961 unique compounds. Unfortunately, the provided NIST MS search software is only available for windows operating systems and does not provide a relatively simple integration into third party software. Currently, our group uses an in-house reference library containing around 400 reference spectra, which covers the most important parts of metabolism, but is certainly not extensive enough. For that reason we acquired the NIST08 reference library in ASCII text format to utilize it within the MetaboliteDetector software package [Hiller et al., 2009]. However, within MetaboliteDetector the target spectrum is compared to *all* spectra within the reference library for identification. If the computational time for

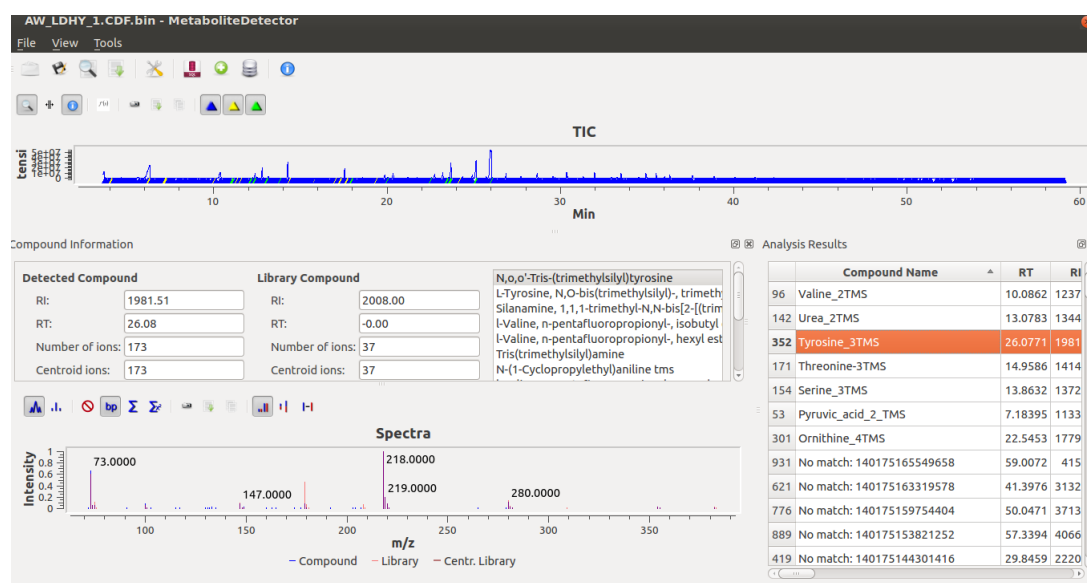


FIGURE 2.3: MetaboliteDetector library search

one spectrum comparison is 1ms then one identification using the NIST library within

MetaboliteDetector would take around 240s. For a typical metabolomics experiment with more than 500 metabolites the computation time will be around 33h which makes it inconvenient for most practical purposes. One way to reduce the computational time of one identification is to compare the target spectrum only to a subset of the reference library. This subset can be generated with the ICBM algorithm. The overall steps are as follows:

- Preprocessing
 - Determine isotope clusters of all spectra within the reference library
 - Store m/z of monoisotopic peak of the five most abundant isotope clusters
- Generate list of possible matches
 - Determine isotope clusters of target spectrum
 - Find all spectra within the reference library that have at least x (e.g. 5) of the isotope clusters (matching m/z value of the monoisotopic peak)
- Find best match
 - Calculate spectrum similarity score to each spectrum in the list of possible matches

The total number of comparisons is reduced significantly by using the strategy described above. To take advantage of the library search functionality, I implemented this feature as well as the ICBM algorithm in the current version of the MetaboliteDetector. Currently, a mass spectral library in NIST format can be imported to an SQLite database. This database can then be queried for selected compounds within MetaboliteDetector's graphical user interface. Figure 2.3 depicts the search result for tyrosine 3-TMS with the NIST08 as the underlying reference library.

The advantage of utilizing a big reference library such as the NIST library became apparent for a discovery made recently in our institute. In a non targeted metabolomics experiment we found highly elevated levels of itaconic acid in LPS treated macrophages compared to non-treated macrophages. Since itaconic acid was not known to be produced by mammalian cells, it was not present in our reference library. Nevertheless, it popped up as one of the unidentified compounds in our non-targeted analysis. In fact, it was one of the most significantly changed metabolites [Michelucci et al., 2013]. The application of the above described ICBM library search functionality helped to identify this compound as itaconic acid. (Figure 2.4). For itaconic acid the m/z values of the monoisotopic peaks of the 5 most abundant isotope clusters are: 147 m/z , 73 m/z , 215 m/z , 259 m/z , and 97 m/z . The NIST08 spectral library contains roughly 200,000

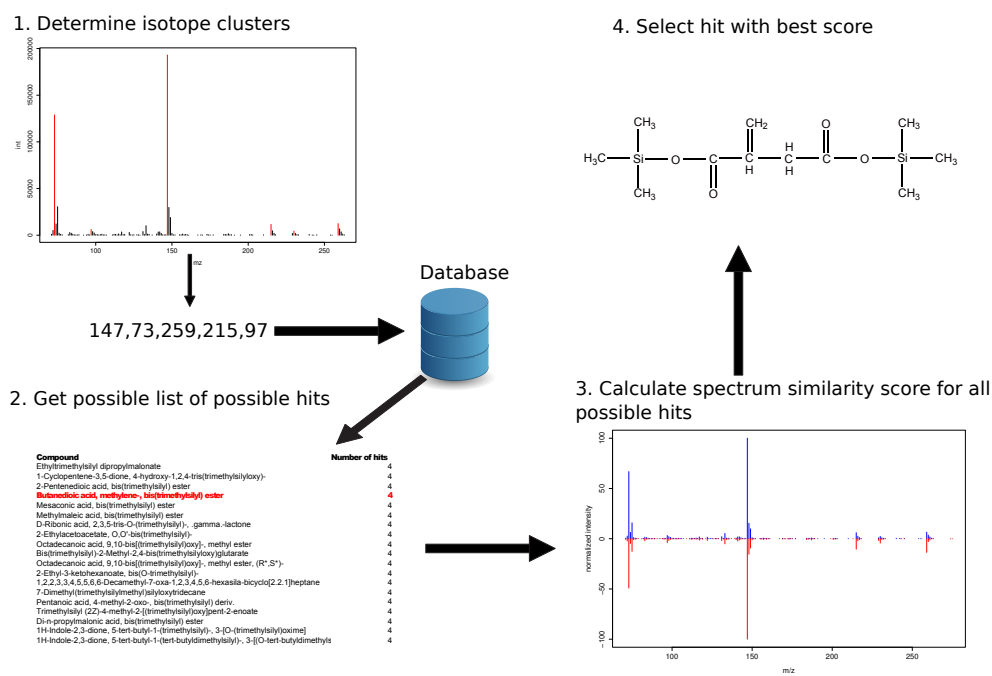


FIGURE 2.4: **Overview of ICBM library search.** 1) Determination of isotope clusters of the measured spectrum. In the case of itaconic acid the m/z values of the monoisotopic peak of the five most abundant isotope clusters are: 147 m/z , 73 m/z , 215 m/z , 259 m/z , and 97 m/z . 2) The library is queried for the m/z values of the five most abundant isotope cluster determined in the previous step. A list of possible hits is generated, including all compounds that have at least 3 of the queried isotope clusters. 3) A spectrum similarity score is calculated for all compounds in the possible hit list. 4) The entry with the highest spectrum similarity score is assigned to the measured compound

EI mass spectra. The list of possible hits, however, generated with the above stated isotope clusters only contains 19 mass spectra (see Figure 2.4). Although the isotope clusters at 147 m/z and 73 m/z originate from the derivatization reagent and usually do not carry any discriminating information, in case of big reference libraries containing mass spectra of different derivatization reagents, they help to highlight the correct compounds. It should be noted that the itaconic acid could have been also identified with a different library search program (e.g. with NIST MS search software). Nevertheless, the integration of the ICBM library search functionality eminently improves the usability of MetaboliteDetector.

In summary, the ICBM algorithm improves a non-targeted metabolomics experiment in two ways: First, it improves the alignment of compounds across different chromatograms, which is one of the main bottlenecks of a non-targeted metabolomics analysis. Second, it helps to identify compounds when a reference library is used.

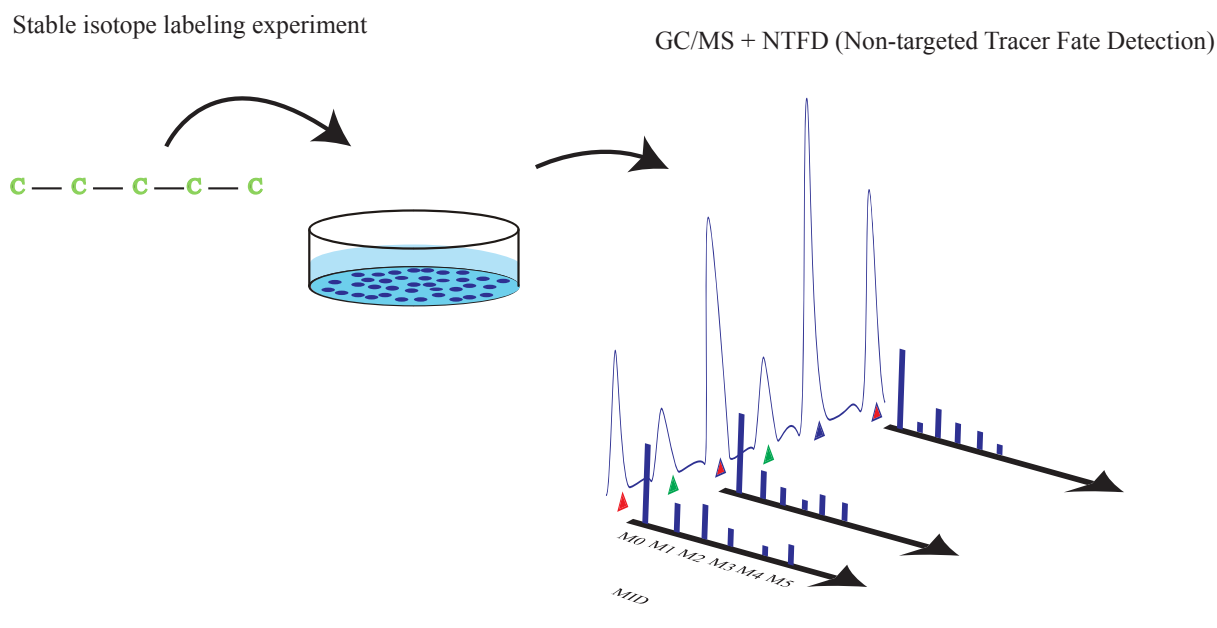


FIGURE 2.5: **Non-targeted tracer fate detection (NTFD)**. Stable-isotope labeled compounds (in this case ^{13}C) are fed to the cell culture. After a GC/MS measurement, the NTFD algorithm automatically detects all labeled compounds and automatically calculates mass isotopomer distributions (MIDs) for all detected compounds. Since NTFD is non-targeted, not only MIDs for known labeled compounds (red triangle) are calculated, but also MIDs for unknown labeled compounds (red triangles with blue frame). Known compounds without label are shown as green triangles, unknown compounds without label as blue triangles.

2.2 Non Targeted Tracer Fate Detection

The discovery of itaconic acid was relatively straightforward, because of its high intracellular abundance in LPS stimulated macrophages compared to unstimulated macrophages. Itaconic acid turned out to be an endpoint of an intracellular flux as a part of the mammalian immune response. In case of metabolic cycles or parallel pathways, however, measuring metabolite concentrations is not sufficient to detect changes in intracellular metabolic fluxes. Moreover, often only pathway intermediates can be measured. As the intracellular concentrations of those pathway intermediates usually do not change significantly, a stable isotope labeling experiment has to be performed to infer metabolic fluxes from the labeling patterns of these intermediates (see Figure 1.11). Besides the determination of intracellular metabolic fluxes, a stable isotope labeling experiment allows one to follow the fate of a specific stable isotope labeled substrate within a given system. Since cellular metabolism is highly complex and not yet fully understood, the advantage of following the distribution of labeling in the metabolic network in a non-targeted manner is clear. For example, disease specific alterations of metabolism may differ from biochemical knowledge derived from textbooks or databases. One method for the non-targeted detection of stable isotope labeled compounds was described by Hiller

[Hiller et al., 2010] . In this section I will describe the implementation of this algorithm in a stand alone software that allows to detect *all* stable isotope labeled compounds downstream from a given substrate (Figure 2.5).

NTFD can be used as a discovery tool that can probe cellular metabolism in a non-targeted way. Usually, the correct calculation of MIDs is a highly targeted process and

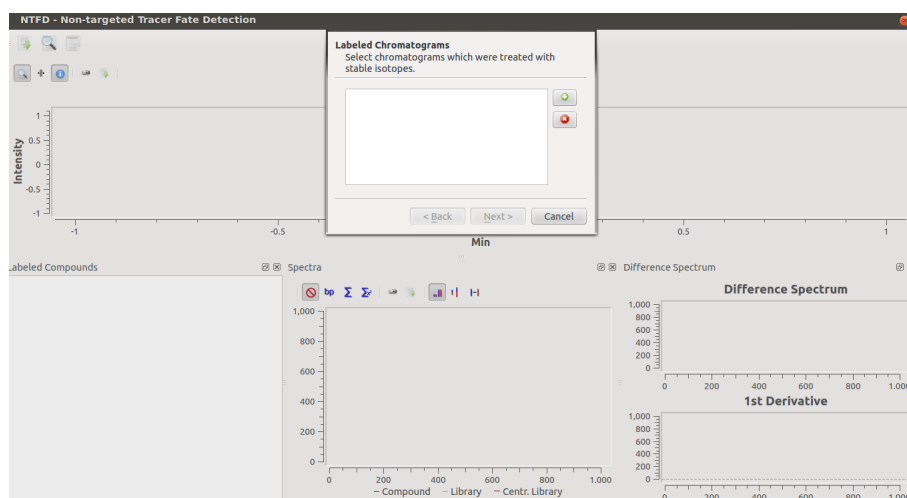


FIGURE 2.6: **NTFD graphical user interface.** The NTFD program provides an easy to use graphical user interface which allows the user to import GC/MS data in netCDF format. The NTFD program then performs automatically the compound detection and detects all labeled compounds within the provided chromatograms. Finally, the MIDs for those labeled compounds are calculated.

requires detailed information about the compounds of interest prior to the analysis. This information includes either the unlabeled reference spectrum or the chemical formula for the fragment ion of interest. In theory, one could investigate every potentially labeled mass spectrum and compare it to the corresponding unlabeled mass spectrum in order to detect all labeled compounds. However, this is a time consuming process and in case of a low percentage of enrichment, almost impossible to catch by manual inspection. For that reason, we developed the NTFD software package with a graphical user interface (Figure 2.6) and made it publicly available for the metabolomics community at <http://ntfd.mit.edu>. The NTFD program can import GC/MS data in netCDF format and performs the following steps automatically:

- Compound detection and chromatographic deconvolution
- Chromatogram alignment
- Detection of labeled compounds
- Calculation of MIDs for labeled compounds
- Compound identification with a reference library

- Export of result in a tab separated format

2.3 Metabolome Of The Neuronal Cell Line LUHMES

Dopaminergic neurons are primarily found in the substantia nigra pars compacta of the midbrain. Although they are few in numbers (usually less than 1% of the total number of brain neurons), they play an important role in the control of multiple brain functions such as voluntary movement and behavioral processes [Chinta and Andersen, 2005] and constitute the major source of dopamine in the mammalian central nervous system. The progressive degeneration of dopaminergic neurons is the major hallmark of Parkinson's Disease (PD). PD is one of the most common neurological disorders, affecting around 1–2% of the over 55 years old population, with differences in gender and increased prevalence with ageing. To date, several different processes including inflammation, oxidative stress and mitochondrial dysfunction have been hypothesized to be the cause of this loss of dopaminergic neurons. As dopamine metabolism itself can be a source of oxidative stress [Meiser et al., 2013], I analyzed the metabolome of the human dopaminergic neuronal cell line LUHMES [Scholz et al., 2011]. LUHMES are

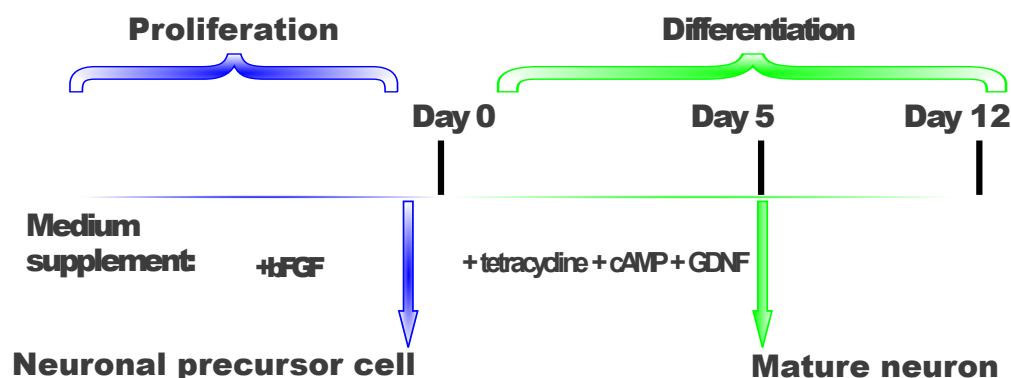


FIGURE 2.7: **LUHMES differentiation.** Neuronal precursor cells can be proliferated by adding cytokine basic fibroblast growth factor (bFGF) as a medium supplement. In absence of bFGF these neuronal precursor cells can be differentiated to mature neurons by adding tetracycline, dibutyryl cAMP (cAMP), and glial cell derived neurotrophic factor (GDNF) to the medium.

human mesencephalic cells conditionally immortalized with a v-myc retroviral vector to ensure continuous proliferation. Inactivation of this vector with tetracycline allows differentiation into mature neurons within 5 to 12 days (Figure 2.7).

2.3.1 Oxygen Level

It has been shown previously that differentiation of neuronal precursor cells to dopaminergic neurons is enhanced under low oxygen conditions [Studer et al., 2000]. Since the *in vivo* oxygen level in mammalian brains is as low as 1% to 5% [Studer et al., 2000],

I intended to test whether the metabolome of proliferating vs differentiated LUHMES cells is affected by the oxygen level. For that reason, I cultured proliferating LUHMES cells at a low oxygen level (5%) and at standard cell culture conditions (20%). Likewise, I differentiated the LUHMES cells at 5% and 20% oxygen. After nine days I extracted the intracellular metabolites and subsequently measured them with GC/MS (analytical details can be found in ??).

After GC/MS measurement, I applied the ICBM algorithm to match compounds across the four conditions. Figure 2.8 depicts the principal component analysis (PCA) of this experiment. The metabolomes of proliferating and differentiated LUHMES cells are well

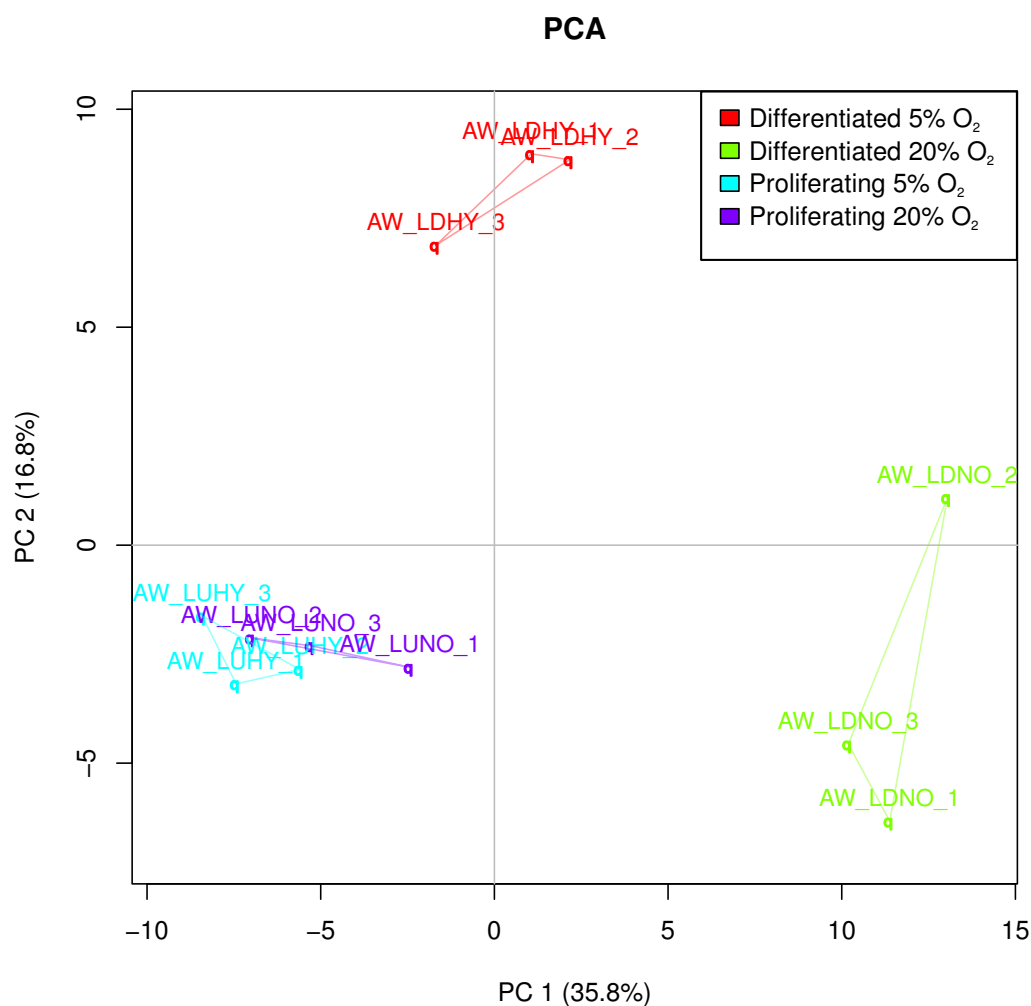


FIGURE 2.8: **PCA of differentiated and proliferating LUHMES cells at 5% and 20% oxygen**

separated, so are the metabolomes of differentiated LUHMES cells at 5% and 20% oxygen. However, the metabolome of proliferating LUHMES cells cannot be distinguished based on the two oxygen levels when all four conditions are considered.

At the metabolome level, there is a difference between LUHMES cells differentiated at 5% oxygen and LUHMES cells differentiated at 20% oxygen. To further study the differences in the metabolome of proliferating and differentiated LUHMES cells at different oxygen levels, I performed an analysis of variance (ANOVA) to detect all significantly (Bonferroni corrected p-value < 0.0003) changed metabolites. Overall, I was able to detect 5 significantly changed metabolites. The corresponding bar plots are depicted in Figure 2.9. Interestingly, lactic acid levels are higher in differentiated LUHMES cell compared to proliferating LUHMES cells. This is in contrast to the observation that proliferating cells mainly rely on aerobic glycolysis to produce adenosine 5-triphosphate (ATP) [Vander Heiden et al., 2009]. However, I am presenting intracellular metabolite levels here and lactic acid is usually excreted from the cell. The lactic acid levels in differentiated LUHMES cells at 5% oxygen are higher compared to differentiated LUHMES cells at 20% oxygen, because of the absence of O₂ to fuel mitochondrial oxidative phosphorylation. An interesting difference between differentiated LUHMES cells at the two oxygen levels is the increased abundance of γ -aminobutyric acid (GABA) at 5% oxygen. GABA is the major inhibitory neurotransmitter in the mammalian central nervous system. Recently it has been shown that dopaminergic neurons can release GABA via the vesicular monoamine transporter VMAT2, which is also the vesicular transporter for dopamine [Tritsch et al., 2012]. This underlines the fact that for dopaminergic cell culture models a low oxygen level should be applied to better reflect conditions in the brain.

In conclusion, this experiment showed that the metabolome of differentiated LUHMES cells at 5% and 20% oxygen is clearly different. These differences do not only originate from metabolites that have been shown before to be affected by hypoxic conditions, such as lactic acid or citric acid [Vander Heiden et al., 2009]. For example, GABA is not directly linked to increased aerobic glycolysis. Nevertheless, it is increased in differentiated LUHMES cell at 5% oxygen. This suggests that indeed the differentiation of dopaminergic neurons is enhanced at low oxygen levels.

2.3.2 Dopamine Metabolism

Although LUHMES should be dopamine producing cells [Scholz et al., 2011], I was not able to detect dopamine in any of the four conditions. To estimate our analytical sensitivity for dopamine by GC/MS, I determined the limit of detection with a dilution series of dopamine standards, starting with 1.5mg/mL. As a result I calculated that we are able to detect dopamine concentrations down to the nanomolar range in full scan mode and down to the picomolar range in single ion mode. To further test if the applied extraction protocol is suitable for dopamine detection, we extracted and measured

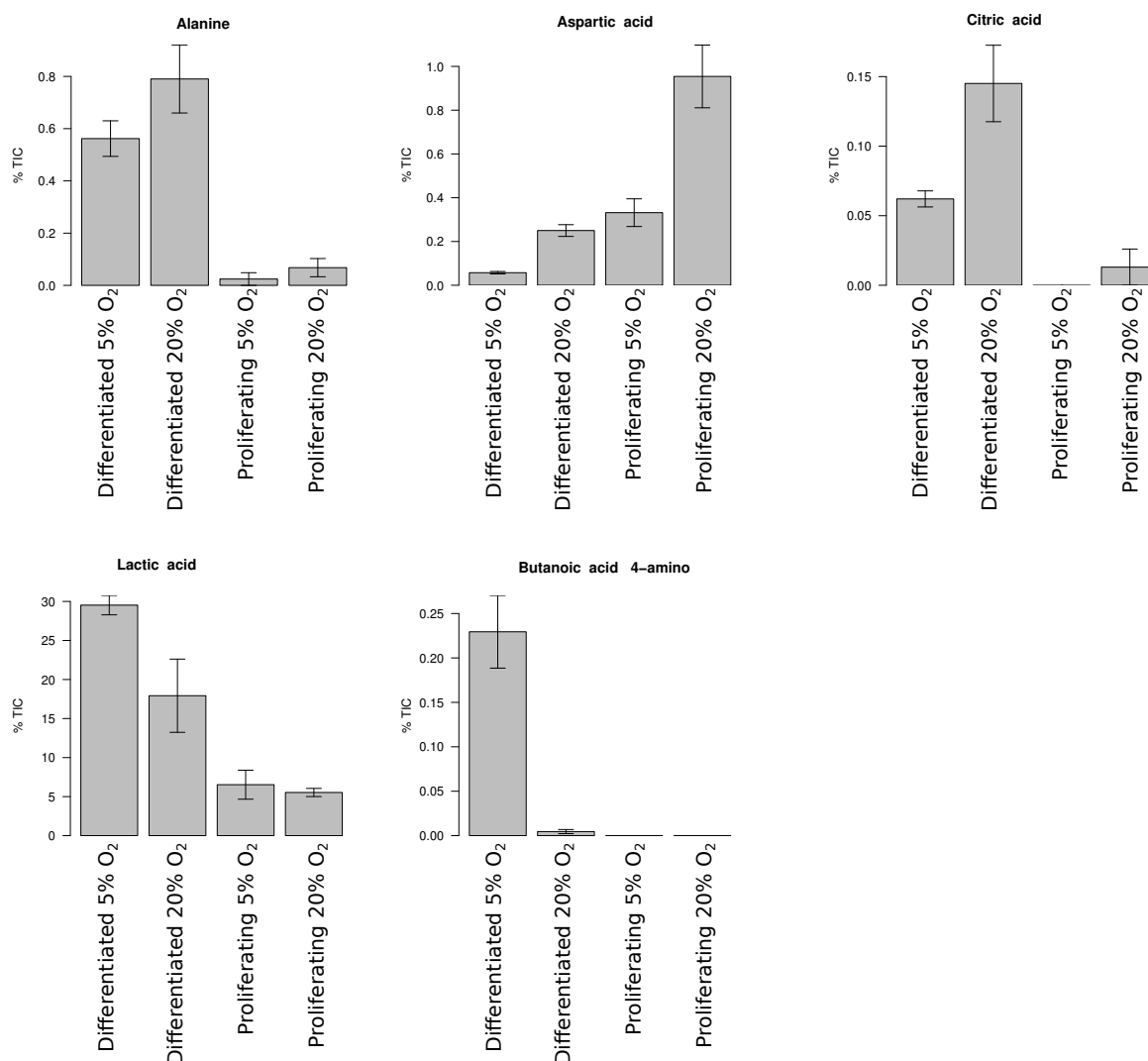


FIGURE 2.9: Bar plots for selected metabolites of proliferating and differentiated LUHMES cells at 5% and 20% oxygen.

mouse brain extracts. Although a mouse brain contains only a few thousand dopaminergic neurons [German and Manaye, 1993], a clear dopamine peak was detectable (Figure 2.10). Together with Dr Johannes Meiser, we tested if the most important enzyme of dopamine synthesis, tyrosine hydroxylase (TH), is present in differentiated LUHMES cells. The Western blot for TH depicted in Figure 2.11 was performed within our group by Dr. Johannes Meiser. Based on the Western Blot, TH protein is present in differentiated LUHMES cells at 2% and 20% oxygen, but not in proliferating LUHMES cells. It is important to note that the TH protein abundance is much higher at 2% oxygen compared to 20% oxygen. This result further endorses the fact that neuronal differentiation is enhanced under low oxygen conditions. In conclusion, we have a highly sensitive method for dopamine detection (picomolar), we are able to detect dopamine in mouse brain extracts, and TH is present in differentiated LUHMES cells, but I was not able to

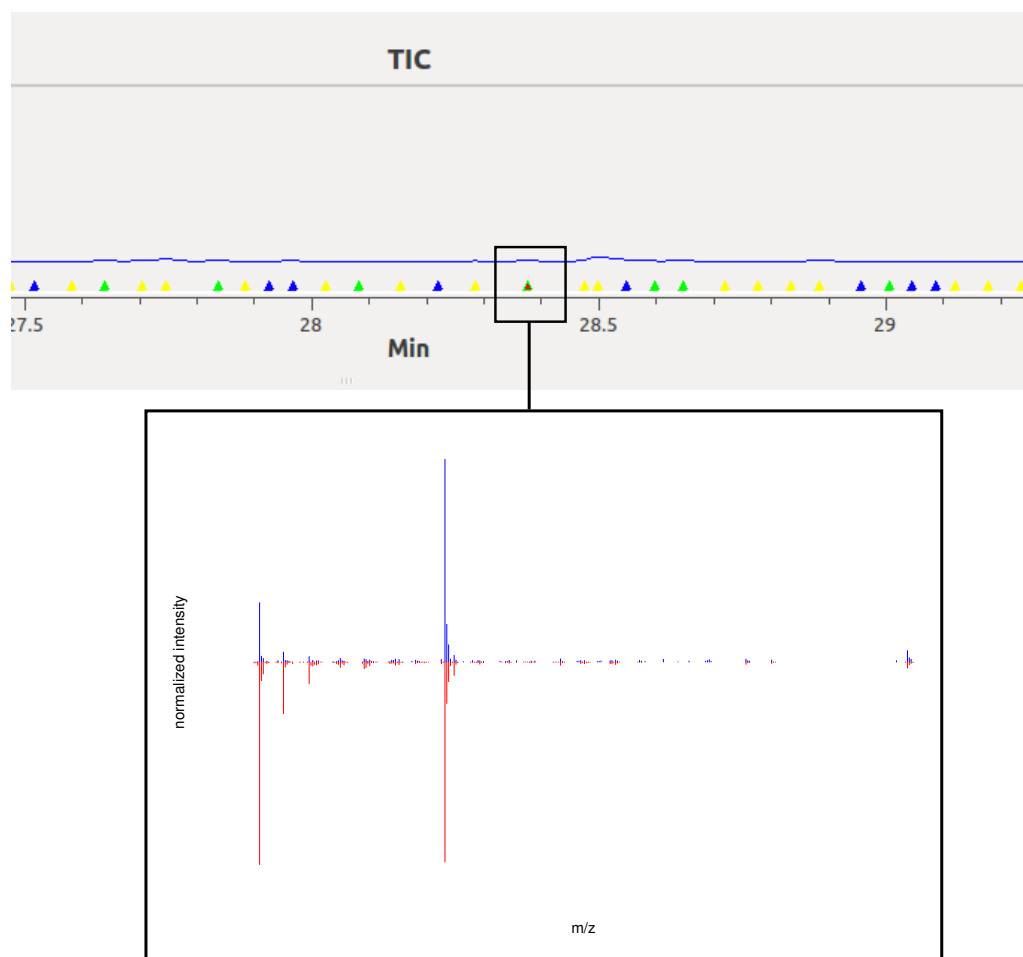


FIGURE 2.10: **GC/MS measurement of mouse midbrain extract.** A part of the total ion chromatogram (TIC) is shown in the upper part. Dopamine 4TMS elutes at 28.38 minutes. The lower part shows the mass spectrum of Dopamine 4TMS. The measured spectrum is shown in blue in the positive direction and the library spectrum is shown in red in the negative direction.

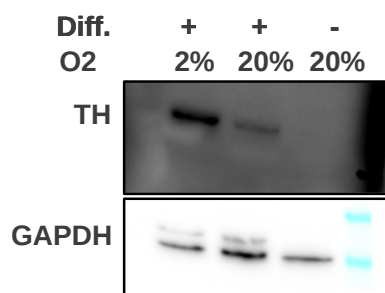


FIGURE 2.11: **Tyrosine hydroxylase abundance.** This Western blot was kindly provided by Dr. Johannes Meiser.

detect dopamine in differentiated LUHMES cells. This result suggests that the presence of TH alone is not sufficient to classify cells as dopamine producing cells and underlines the importance of metabolomics to study cellular phenotypes.

Chapter 3

Targeted Metabolomics Methodologies

This chapter covers a manuscript about a methodology for the determination of chemical formulas and retained atoms for mass spectral fragment ions. The main results are summarized with respective cross references to the manuscript given. For methodological details, please consult the corresponding manuscript [[Wegner et al., 2014](#)].

Wegner, A.; Weindl, D.; Jäger, C.; Sapcariu, S. C.; Dong, X.;
Stephanopoulos, G.; Hiller, K.

Analytical chemistry

3.1 Fragment Formula Calculator

As stated in section [1.1.3](#), non-targeted methodologies are of great importance, but they cannot replace targeted approaches. For example, MIDs are of high importance for stable isotope labeling experiments and can be calculated in a non-targeted way with the NTFD algorithm. However, to make biological sense out of MIDs, detailed information about the underlying fragment ions are necessary. Specifically, the structural formulas for fragments ions are essential to pull out biological information of MIDs. This information can be obtained only in a targeted way and requires the structural formula of the molecular (Figure [3.1](#)). For example, ^{13}C -MFA relies on exact knowledge of the position of the label in order to determine intracellular fluxes from MIDs using a nonlinear least-squares parameter estimation approach (see Section [1.3](#)). In this process, MIDs obtained from a stable isotope labeling experiment help to constrain the fluxes in

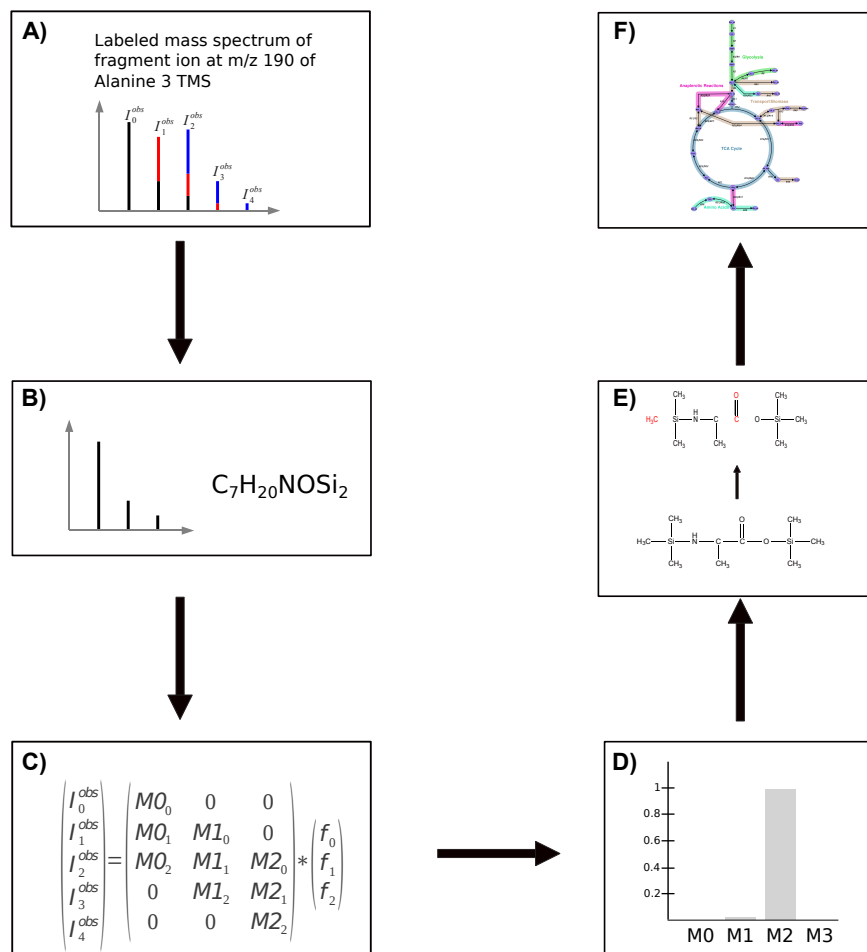


FIGURE 3.1: **Importance of targeted methodologies.** A) The labeled mass spectrum for the fragment ion at m/z 190 of Alanine is shown. B) In order to calculate MIDs, either the chemical formula or the unlabeled reference spectra must be available. C) Set-up of the linear equation system. Note that this is also possible in a non-targeted methodology when the unlabeled reference spectrum is applied. D) MIDs can be calculated in a target and non-targeted way. E) The retained carbon atoms can be only determined if the structural formula of the molecular ion is available. Hence this can only be done in a targeted way. F) The calculated MIDs in combination with the retained carbon atoms can be applied in ^{13}C -MFA.

a given system. In particular, MIDs for fragment ions containing different carbon atoms are of high interest, since they can carry different flux information. Therefore it is crucial to identify the structural formulas for mass spectral fragment ions. Since ^{13}C -MFA is based on carbon labeling, it is sufficient to identify the retained carbon atoms. However, the process of assigning a chemical formula and retained atoms to mass spectral ions is non-trivial and time-consuming, even for an expert. For that reason, I developed an algorithm that can determine the chemical formulas and the retained atoms for mass spectral fragment ions. Generally, there are two ways to determine the retained atoms of a fragment ion: a rule-based *in silico* prediction or a combinatorial approach based on a systematic bond cleavage. Rule-based algorithms rely on fragmentation

mechanisms derived from molecules where the fragmentation is known, assuming that similar structures will fragment the same way. However, small changes in structure can lead to a significantly different fragmentation mechanism. Furthermore, the rule-based approach fails for molecules where no similar fragmentation mechanism is known. Here, I present a method to determine the chemical formulas and the retained atoms for mass spectral fragment ions without *a priori* knowledge about the fragmentation mechanisms, taking advantage of the combinatorial aspect of the problem. For that, I will apply the molecule's graph-theoretical representation to model the fragmentation.

3.1.1 Algorithm

I model a molecule as an undirected, connected and labeled graph $G = (V, E, f_{VA}, f_{VB}, f_{VC}, f_{ED})$, where V is the set of vertices corresponding to the atoms and E is the set of undirected edges corresponding to the bonds between the atoms. The function $f_{VA} : V \rightarrow A$ assigns each atom an element (e.g. carbon, hydrogen, etc.), $f_{VB} : V \rightarrow B$ assigns each atom an index and $f_{VC} : V \rightarrow C$ assigns each atom the atomic mass according to the chemical element. The function $f_{ED} : V \times V \rightarrow D$ assigns each bond an order (single, double, triple). The mass of the molecular ion corresponds to the sum of the masses of all vertices:

$$W(G) = \sum_{v \in V} f_{VC}(v) \quad (3.1)$$

The underlying idea of this algorithm is that the fragmentation process usually only breaks a few bonds within the molecule. This can be simulated by removing a defined number of edges within the molecular graph. In terms of graph theory this means to induce a cut of a certain size in the graph. This can leave the graph G disconnected. The resulting connected components $C = \{C_1, \dots, C_n\}$ of the subgraph H each have a molecular mass:

$$W(C_i) = \sum_{v \in V(C_i)} f_{VC}(v) \quad (3.2)$$

Since the mass m of the fragment ion is determined by mass-spectrometry, the chemical formula of this fragment ion corresponds to a combination of connected components of H which molecular masses $W(C_i)$ sum up to m . Figure 3.2 illustrates this process. The resulting subgraph (representing the chemical composition), which can be composed of several connected components, does not necessarily represent the chemical structure, because the formation of new bonds (e.g. fragmentation rule 4 described in Section 1.1.1.2) is not modeled. However, the number and position of atoms of the intact compound retained in this fragment ion is uncovered.

So far, I have relied on the assumption that the correct edges are deleted from the graph. There are two unknowns, the number and the position of edges to be deleted. To define

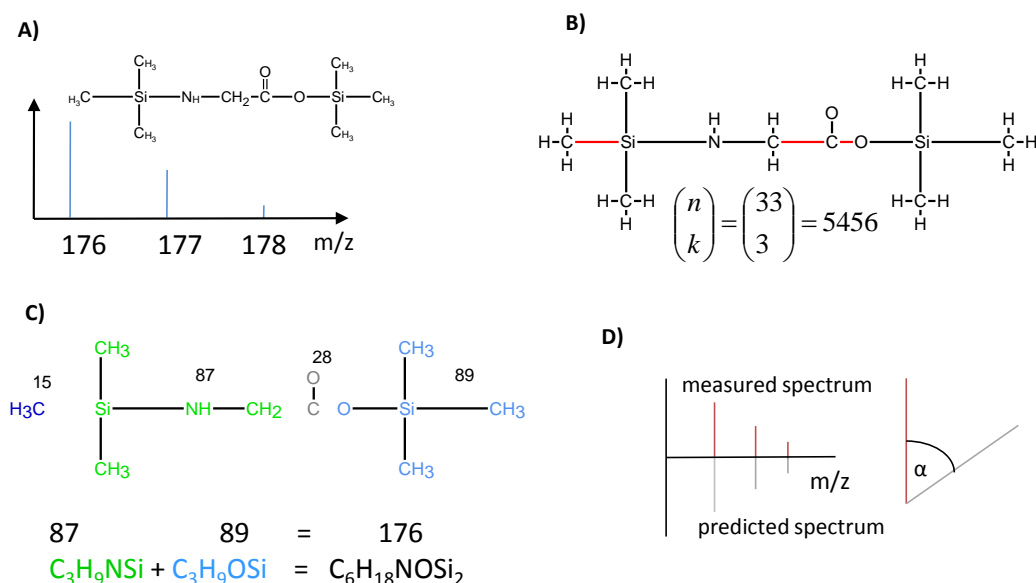


FIGURE 3.2: Overview of FFC algorithm. (A) As input FFC needs the 2D structure of the compound together with the mass spectrum of the ion of interest. In this example, I present the molecule *N,O*-bis-(trimethylsilyl)-glycine (219 Da) and the fragment ion at mass 176. (B) The 2D structure is first converted into a molecular graph. The graph contains 34 vertices and 33 edges. Then all combinations of edge sets of a certain size (in this case 3) are consecutively deleted from the graph, resulting in 5456 disconnected graphs, one for each edge set deleted. The number of resulting subgraphs can be calculated with the binomial coefficient where n corresponds to the number of edges and k corresponds to the cut size (Equation 3.3). For simplification, only the edge set leading to the correct fragmentation is shown here. (C) For each disconnected graph the connected components are determined. For every combination of connected components where the molecular masses sum up to the mass of the fragment ion, the atoms of these components are combined to build up a candidate formula. In this example, the connected components shown in green and light blue with the masses 87 and 89 sum up to the target mass of 176. The candidate formula is then $\text{C}_6\text{H}_{18}\text{NO}_2\text{Si}_2$, which is indeed the correct formula for this fragment ion. In addition to the chemical formula, the algorithm also yields positional information about the fate of specific atoms. For example, the carboxyl carbon of the original glycine molecule is lost in this fragment ion. (D) Based on the candidate formula the theoretical mass spectrum is predicted and a spectrum similarity score to the measured spectrum based on the dot product is calculated. This is of special importance if more than one sum formula can be derived for the target mass.

the minimal number of edges to delete from the graph (cut size), necessary to model the fragmentation, it is mandatory to take the fragmentation rules (as stated in Section 1.1.1.2) into consideration. Fragmentation types 1-3 cleave one bond without forming new σ -bonds, 4 and 5 cleave one bond while forming a new one, 6 cleaves two bonds while forming a new one. Therefore, to describe an α -cleavage or a σ -ionization, clearly a cut size of one is sufficient. To simulate a simple elimination or a rearrangement, which is equivalent to deleting one edge in the graph, a cut size of one is also necessary. For the combination of a more complex rearrangement and an α -cleavage, a cut size of three is necessary. To capture both the single and the combined fragmentations, the algorithm is designed to work with a defined maximum cut size. The cut size starts at one and subsequently increases until it reaches the defined maximum cut size.

One way to find the correct edges to delete from the graph is to select those edges that are most likely to break. For example, low-energy bonds can be assumed to break more easily. Although this is correct, additional rules are needed to describe rearrangements. Another more straightforward way is to delete all possible combinations of edges of a certain cut size. Certainly this includes the correct edges, but at the same time increases the number of possible results enormously. If the number of edges is given by n and the cut size by k , then the number of k distinct elements of n is given by the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \quad (3.3)$$

For example, the graph of the molecule *N,O*-bis-(trimethylsilyl)-glycine with the molecular formula $C_8H_{21}NO_2Si_2$ has 33 edges. The number of possible distinct edge sets to delete for a cut size of 3 is then 5456.

To find the correct edges, the resulting fragment formulas for each of these possibilities have to be ranked according to a score. At best, this score is linked to the measured mass spectrum. One elegant way to do so is to predict the theoretical mass spectrum of the determined fragment formula and calculate a spectrum similarity score to the measured mass spectrum of this fragment ion. A mass spectrum can be theoretically predicted by using the natural stable isotopic distribution of elements and statistical theory [Fernandez et al., 1996]. For elements that only have one naturally occurring stable isotope of significant abundance, the distribution of isotopes can be predicted by a binomial distribution:

$$m_i = \frac{n!}{i! \cdot (n-i)!} \cdot p_0^{n-i} \cdot p_1^i \quad (3.4)$$

where n is the total number of atoms, i the number of atoms containing the heavier isotope (e.g. ^{13}C), p_0 the natural abundance of the lighter isotope (e.g. $p(^{12}C)=0.989$) and p_1 the natural abundance of the heavier isotope (e.g. $p(^{13}C)=0.01$). In case an element has several natural occurring isotopes the distribution of those isotopes within

a molecule can be predicted by a multinomial distribution:

$$m_i = \frac{n!}{a_1! \cdot a_2! \cdot \dots \cdot a_k!} \cdot p_0^{a_0} \cdot p_1^{a_1} \cdot \dots \cdot p_k^{a_k} \quad (3.5)$$

where n is the total number of atoms, a_0 to a_k the number of atoms containing the respective isotope and p_0 to p_k the natural abundances of those isotopes.

3.1.1.1 Reducing Algorithmic Complexity

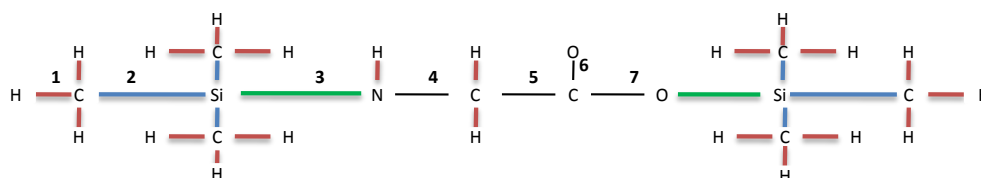


FIGURE 3.3: **Graph representation of *N,O*-bis-(trimethylsilyl)-glycine.** The graph contains 33 edges. For a cut size of three the number of distinct edge pairs to delete is 5456. To reduce the number of distinct edge pairs, non backbone edges (edges that are not connected to at least one backbone atom) are grouped based on their loss pattern. For example, edges shown in red are grouped together because their removal leads to the loss of one hydrogen. The group of edges shown in blue lead to the loss of a methyl group when one of these edges is removed. The group of edges shown in green lead to the loss of a TMS group when one of these edges is removed. After reduction to relevant backbone edges, the graph now contains only 7 distinct edge groups (as illustrated by the numbers above the edges) which reduces the number of distinct edge sets of size 3 from 5456 to 35.

For GC/MS, compounds are usually derivatized prior to analysis. For example, hydrogens in polar functional groups can be replaced with a trimethylsilyl (TMS) or *tert*-butyldimethylsilyl (TBDMS) group (see Section 1.1.1.3). This makes compounds more volatile and less reactive, but at the same time increases the computational complexity of finding the correct chemical formula of a fragment ion. In case of stable isotope labeling experiments, the interest lies normally only in labeling patterns for atoms of the original (underivatized) molecule. As a consequence, the information obtained from the loss of atoms originating from the derivatization agent used is often redundant. For example, when TMS derivatization is used, a $[M-15]^+$ fragment is often present in the mass spectrum, originating from the loss of a methyl group from the derivatized part of the molecule. Depending on the number of TMS groups within the molecule, there are several possibilities for the position of the lost methyl group. Concerning the calculation of chemical formulas, however, the position of this methyl group is not relevant and computational time can thus be saved. For that reason, we divide the molecular graph into atoms belonging to the original molecule (backbone atoms) and atoms originating from

the derivatization agent used. Subsequently, non backbone edges (edges that are not connected to at least one backbone atom) are grouped based on the atoms that would be lost if this edge is deleted (Figure 3.3). For example, all edges are grouped together where their removal would lead to the loss of one hydrogen. This reduces the number of distinct edges significantly, thereby decreasing the combinatorial complexity for the problem of finding the correct chemical formula. Additionally, this allows the user to follow the fate of specific atoms in the molecular ion by selecting them as backbone atoms.

Another advantage which makes the proposed algorithm capable of modeling rearrangements is the use of connected components. Fragment ions resulting from a rearrangement reaction are often composed of two or more disjoint substructures of the molecular ion. Identifying these substructures is computationally challenging, as their number grows enormously with the number of atoms. However, in our algorithm the number of these substructures is limited by the number of connected components within the molecular graph, making the proposed algorithm also applicable for larger molecules.

3.1.1.2 Constraining The Result Set

One major advantage of the FFC program compared to commercial softwares such as ACD/MS Fragmenter or Mass Frontier is that it is able to incorporate stable isotope labeled spectra in the analysis. In most cases multiple candidate formulas are available for one fragment ion and it is not immediately clear which of those formulas is the correct one. In these cases, a stable isotope labeled spectra of the compound of interest can help to remove wrong candidate formulas from the the result set. For that, the FFC program assigns each atom within the molecule a binary state: 1 means this atom is present in this fragment ion and 0 means it is cleaved off. Consequently, each candidate formulas is described by a set of bits depending on the atoms present in this candidate formula. In order to include the labeled spectra in the analysis, the FFC program uses a second bit (labeled bit set) set to describe the applied stable isotope tracer. This bit set describes the labeling state of each atom within the molecule: 1 means this atom is labeled and 0 means the atom is unlabeled. For example, $^{13}\text{C}_3$ β -Alanine has 3 labeled carbon atoms. This means the bits corresponding to these three carbon atoms would be set to 1 and the bits corresponding to the remaining atoms to 0 (Figure 3.4). Subsequently, each candidate formula's bit set is combined with the labeled bit set through a logical conjunction. The resulting bit set defines exactly how many labeled atoms are present in this candidate formula and can be easily calculated by counting the number of 1's in this bit set. The number of labeled atoms can then be compared to the measured labeled spectrum. For that, the FFC program automatically calculates the MIDs for this

fragment ion. However, for an accurate MID calculation a correction matrix is needed to solve the linear equation system [Lee et al., 1991]. This correction matrix can either be setup with the chemical formula or the unlabeled reference spectrum of the fragment ion (see Section 1.2.1). Since the chemical formula is the information the FFC program intends to calculate, the software applies the unlabeled reference spectrum to set-up the correction matrix. Finally, all candidate formulas that contradict the MIDs are then automatically excluded from the result set.

The FFC program can calculate MIDs for a wide range of tracers (e.g. ^2H , ^{15}N , ^{13}C or ^{18}O). In case of d_9 -MSTFA, however, the number of deuterium labeled atoms can easily grow above twenty, which can complicate the calculation of MIDs. For that reason, the FFC program does not calculate MIDs for d_9 -MSTFA labeled spectra, but calculates the number of ^2H atoms as follows:

1. Calculate the maximum number of ^2H labeled atoms possible:
$$\text{max}L = \#\text{TMS-groups} \cdot 9$$
2. Find all isotope clusters in the d_9 -MSTFA labeled spectrum within the following range:
m/z value of the monoisotopic peak of the fragment ion of interest (corresponds to 0 ^2H) + $\text{max}L$ (corresponds to the maximum number of ^2H)
3. Calculate a spectrum similarity score (e.g. with the ICBM algorithm) between all isotope clusters determined in the previous step and the fragment ion of interest
4. Select the isotope cluster with the highest spectrum similarity score ($\text{max}S$)
5. Calculate the number of ^2H labeled atoms by subtracting the m/z value of the monoisotopic peak of $\text{max}S$ from m/z value of the monoisotopic peak of the fragment ion of interest

3.1.2 Software Package

Further on I implemented the algorithm in the stand alone software package fragment formula calculator (FFC). Figure 3.5 depicts the graphical user interface of the FFC program. The FFC program is optimized to determine chemical formulas for GC/MS fragment ions.

The input to the program consists of the 2D structure of the compound together with the measured mass spectrum. This data can be loaded directly by the user; a MOL file for the structure and a csv file for the mass spectrum. Additionally, FFC allows the user to import a library in NIST format to an SQLite database which then can

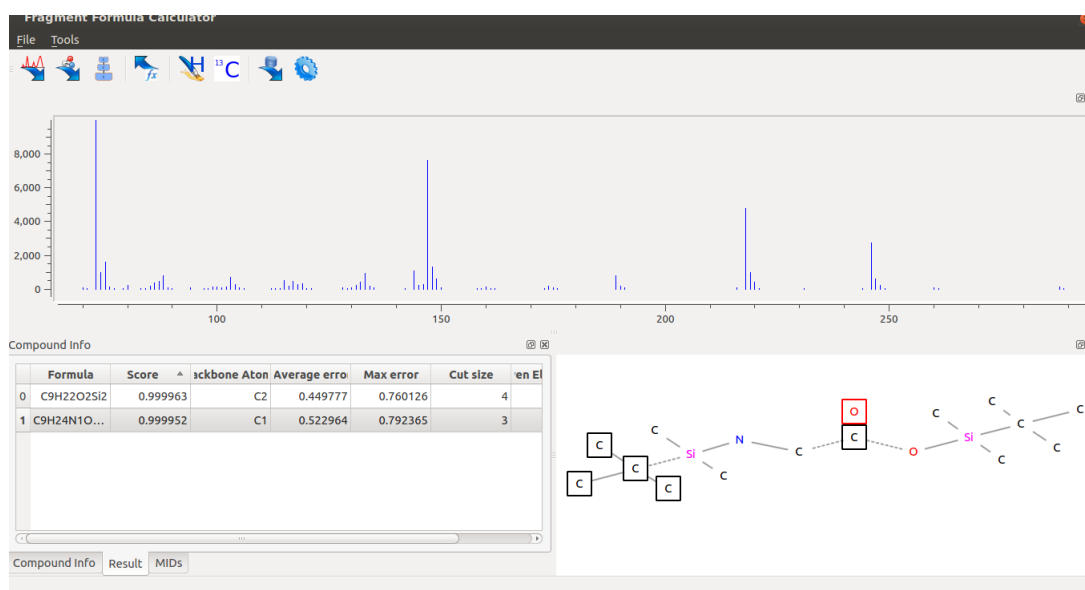


FIGURE 3.5: **FFC graphical user interface.** The window (in the upper part of the main interface) displays the loaded spectrum. The window (in the bottom left of the main interface) contains the compound of interest for fragment calculation shown in the the “Compound Info” tab. The “Result” tab contains the calculated formulas for the compound of interest at the corresponding sized fragment. If a spectrum from a stable isotope labeling experiment was added to the analysis, the “MIDs” tab shows the calculated MIDs for the current selected fragment. The window (in the bottom right of the main interface) displays the chemical structure of the compound of interest. If a result row is selected from the “Result” tab, this window visualizes the fragmentation by highlighting lost atoms with a square and broken bonds with a dashed line.

be queried for a chemical formula or a compound’s name (Figure 3.6). Mass spectral fragment ions are detected automatically based on the mass spectrum [Wegner et al., 2013]. The structure is visualized based on the atom coordinates defined in the MOL file. For clarity, atoms are colored according to the CPK scheme and hydrogens are hidden. Candidate formulas and the corresponding atoms within the molecule can be calculated by pressing the “Start Calculation” button. By default, all atoms within the molecule are considered as backbone atoms, but this can be manually corrected for derivatized compounds. The program generates a list of possible hits which are shown in the “Result tab”. This “Result tab” includes the following entries:

- Chemical formula
- Spectrum similarity score of each predicted mass spectrum to the measured mass spectrum based on the dot product [Stein and Scott, 1994]
- List of backbone atoms present in each candidate formula
- The average and maximum deviation from each measured to the theoretical (predicted) mass spectrum

- The number of broken bonds
- Even electron ion (yes/no)

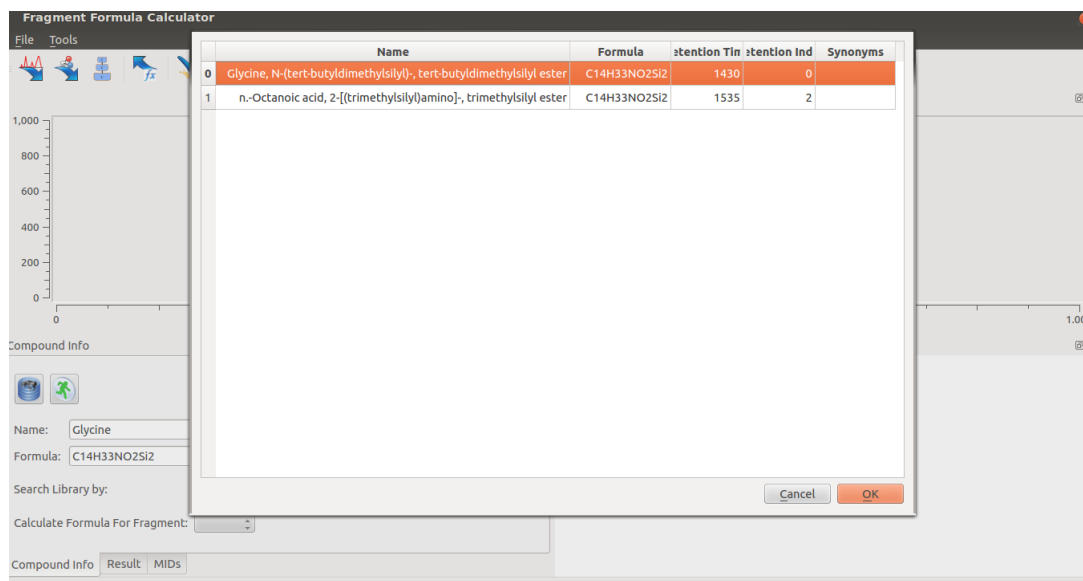


FIGURE 3.6: **FFC library search interface.** The FFC program allows the user to import a library in NIST format to an SQLite database which then can be queried for a chemical formula or a compound's name

3.2 Fragment Formula Repository

As stated in the previous section, the chemical formula in combination with positional information of atoms present in a fragment ion is of high importance to extract biological information out of MIDs. For that reason, I applied the FFC program to determine the chemical formulas and carbon atom composition for a wide range of TMS and TBDMS derivatized compounds of central carbon metabolism.

As mentioned in Section 3.1, the use of stable isotope labeled reference standards can greatly improve the predictive capabilities of the FFC algorithm. However, these labeled reference compounds are very expensive. That is why we generated fully labeled reference spectra in our lab. For that, we used yeast grown on U-¹³C D-glucose as the only carbon source (Figure 3.7). As a result we generated a comprehensive set of fully labeled reference spectra for TMS, d₉-TMS, and TBDMS derivatized compounds (Figure 3.8). Later, I used the unlabeled and fully labeled mass spectra as input for the FFC algorithm and determined the chemical formulas and retained carbon atoms of 160 fragment ions. The chemical formulas for the TMS derivatized compounds can be found in Table 3.1 and for the TBDMS derivatized compounds in Table 3.2.

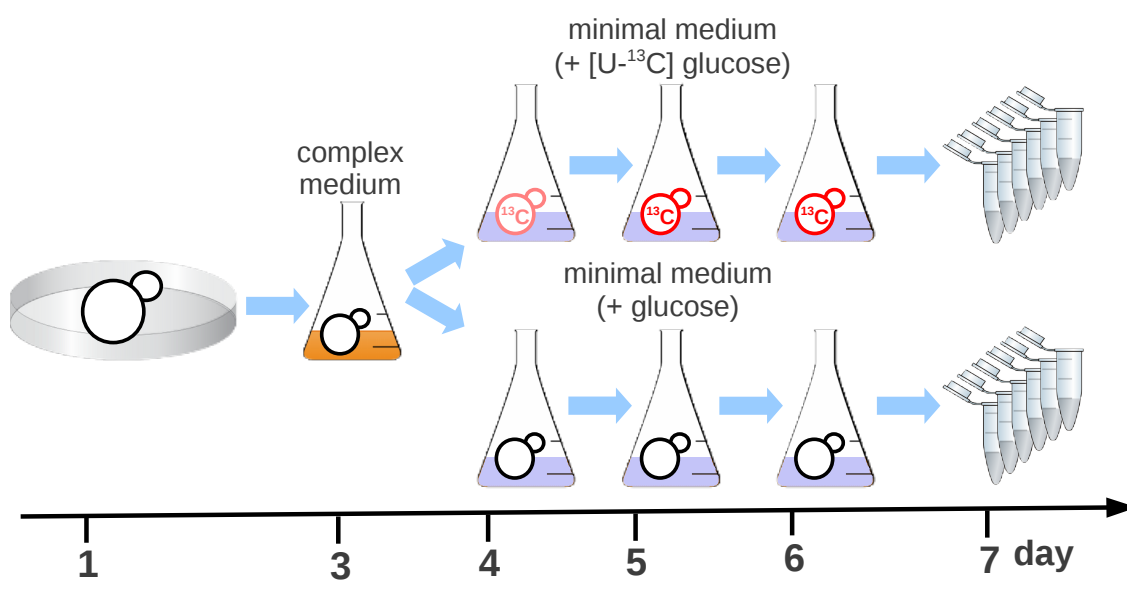


FIGURE 3.7: **Overview yeast culture.** Yeast was cultivated on minimal medium with $U-^{13}C$ D-glucose as the only carbon source. To obtain fully labeled extracts three overnight cultivations were applied before cell harvesting.

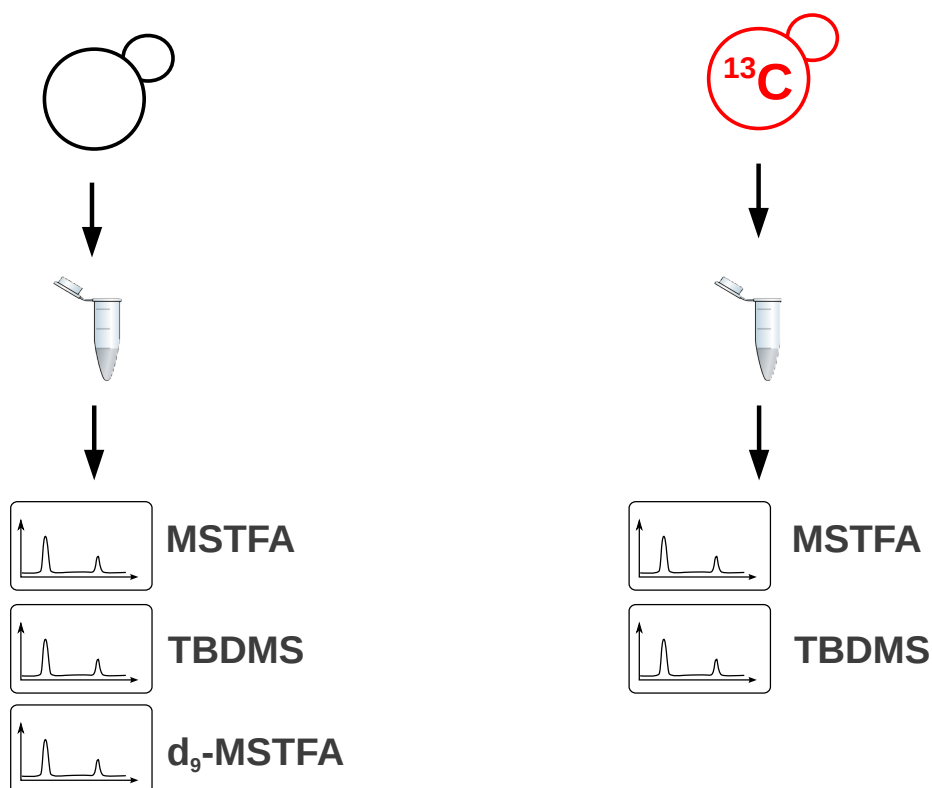


FIGURE 3.8: **Overview yeast measurements.** Labeled and unlabeled yeast extracts were derivatized with MSTFA and TBDMS and measured with GC/MS. The unlabeled extracts were additionally derivatized with d_9 -TMS.

3.2.1 TMS Derivatized Fragment Ions

The retained carbon atoms of all compounds in Table 3.1 can be found in [Wegner et al., 2014]. Atoms that were lost in a fragment ion are shown in red and retained atoms in black.

TABLE 3.1: TMS Derivatized Fragment Ions

Compound	m/z	m/z ¹³ C	m/z d ₉ -TMS	Formula
Adenine 2TMS	279	284	297	C ₁₁ H ₂₁ N ₅ Si ₂
	264	269	279	C ₁₀ H ₁₈ N ₅ Si ₂
	206	211	215	C ₈ H ₁₂ N ₅ Si
Alanine 2TMS	233	-	-	C ₉ H ₂₃ NO ₂ Si ₂
	218	220, 221	233, 236	C ₈ H ₂₀ NO ₂ Si ₂
	190	192	205	C ₇ H ₂₀ NOSi ₂
	116	118	125	C ₅ H ₁₄ NSi
Aspartic acid 2TMS	277	281	295	C ₁₀ H ₂₃ NO ₄ Si ₂
	262	266	277	C ₉ H ₂₀ NO ₄ Si ₂
	234	237	249	C ₈ H ₂₀ NO ₃ Si ₂
	220	222	235	C ₇ H ₁₈ NO ₃ Si ₂
	160	163	169	C ₆ H ₁₄ N ₁ O ₂ Si
Aspartic acid 3TMS	349	354	376	C ₁₃ H ₃₁ NO ₄ Si ₃
	334	338	358	C ₁₂ H ₂₈ NO ₄ Si ₃
	306	309	330	C ₁₁ H ₂₈ NO ₃ Si ₃
	292	294	316	C ₁₀ H ₂₆ NO ₃ Si ₃
	232	235	250	C ₉ H ₂₂ NO ₂ Si ₂
	218	220	236	C ₈ H ₂₀ NO ₂ Si ₂
β -Alanine 3TMS	305	-	-	C ₁₂ H ₃₁ NO ₂ Si ₃
	290	-	314	C ₁₁ H ₂₈ NO ₂ Si ₃
	248	-	272	C ₉ H ₂₆ NOSi ₃
	232	-	250	C ₉ H ₂₂ NO ₂ Si ₂
	174	-	192	C ₇ H ₂₀ NSi
	86	-	92	C ₃ H ₆ OSi
Citric acid 4TMS	480	-	-	C ₁₈ H ₄₀ O ₇ Si ₄
	465	471	498	C ₁₇ H ₃₇ O ₇ Si ₄
	375	381	399	C ₁₄ H ₂₇ O ₆ Si ₃
	363	368	390	C ₁₄ H ₃₁ O ₅ Si ₃
	347	352	371	C ₁₃ H ₂₇ O ₅ Si ₃
	273	278	291	C ₁₁ H ₂₁ O ₄ Si ₂

Continued on next page

Table 3.1 – *Continued from previous page*

Compound	m/z	m/z ¹³ C	m/z d ₉ -TMS	Formula
3-phosphoglycerate 4TMS	474	-	-	C ₁₅ H ₃₉ O ₇ PSi ₄
	459	462	492	C ₁₄ H ₃₆ O ₇ PSi ₄
	387	387	423	C ₁₂ H ₃₆ O ₄ PSi ₄
	357	359	384	C ₁₁ H ₃₀ O ₅ PSi ₃
	315	315	342	C ₉ H ₂₈ O ₄ PSi ₃
	299	299	323	C ₈ H ₂₄ O ₄ PSi ₃
Glycerol-3-phosphate 4TMS	460	-	-	C ₁₅ H ₄₁ O ₆ PSi ₄
	445	448	478	C ₁₄ H ₃₈ O ₆ PSi ₄
	387	387	423	C ₁₂ H ₃₆ O ₄ PSi ₄
	357	359	384	C ₁₁ H ₃₀ O ₅ PSi ₃
	341	343	365	C ₁₀ H ₂₆ O ₅ PSi ₃
	299	299	323	C ₈ H ₂₄ O ₄ PSi ₃
Glutamic acid 3TMS	363	368	390	C ₁₄ H ₃₃ NO ₄ Si ₃
	348	353	372	C ₁₃ H ₃₀ NO ₄ Si ₃
	320	324	344	C ₁₂ H ₃₀ NO ₃ Si ₃
	246	250	264	C ₁₀ H ₂₄ NO ₂ Si ₂
	230	234	245	C ₉ H ₂₀ NO ₂ Si ₂
Glutamine 3TMS	362	367	389	C ₁₄ H ₃₄ N ₂ O ₃ Si ₃
	347	352	371	C ₁₃ H ₃₁ N ₂ O ₃ Si ₃
	273	278	291	C ₁₁ H ₂₅ N ₂ O ₂ Si ₂
	245	249	263	C ₁₀ H ₂₅ N ₂ O ₁ Si ₂
Glycerol 3TMS	308	-	-	C ₁₂ H ₃₂ O ₃ Si ₃
	293	296	317	C ₁₁ H ₂₉ O ₃ Si ₃
	218	221	236	C ₉ H ₂₂ O ₂ Si ₂
	205	207	223	C ₈ H ₂₁ O ₂ Si ₂
Glycine 3TMS	291	293	-	C ₁₁ H ₂₉ NO ₂ Si ₃
	276	278	300	C ₁₀ H ₂₆ NO ₂ Si ₃
	248	249	274	C ₉ H ₂₆ NOSi ₃
	174	175	192	C ₇ H ₂₀ NSi ₂
Isoleucine 2TMS	275	-	-	C ₁₂ H ₂₉ NO ₂ Si ₂
	260	265, 266	275, 278	C ₁₁ H ₂₆ NO ₂ Si ₂
	232	237	247	C ₁₀ H ₂₆ NOSi ₂
	218	220	236	C ₈ H ₂₀ NO ₂ Si ₂
	158	163	167	C ₈ H ₂₀ NSi
Leucine 2TMS	275	-	-	C ₁₂ H ₂₉ NO ₂ Si ₂
	260	265, 266	275, 278	C ₁₁ H ₂₆ NO ₂ Si ₂

Continued on next page

Table 3.1 – *Continued from previous page*

Compound	m/z	m/z ¹³ C	m/z d ₉ -TMS	Formula
	232	237	247	C ₁₀ H ₂₆ NOSi ₂
	218	220	236	C ₈ H ₂₀ NO ₂ Si ₂
	158	163	167	C ₈ H ₂₀ NSi
Lysine 3TMS	362	368	389	C ₁₅ H ₃₈ N ₂ O ₂ Si ₃
	347	353	371	C ₁₄ H ₃₅ N ₂ O ₂ Si ₃
	200	206	209	C ₉ H ₁₈ NO ₂ Si
	174	175	192	C ₇ H ₂₀ NSi ₂
	156	161	165	C ₈ H ₁₈ NSi
Lysine 4TMS	434	440	470	C ₁₈ H ₄₆ N ₂ O ₂ Si ₄
	419	425	452	C ₁₇ H ₄₃ N ₂ O ₂ Si ₄
	391	396	324	C ₁₆ H ₄₃ N ₂ OSi ₄
	317	322	344	C ₁₄ H ₃₇ N ₂ Si ₃
	174	175	192	C ₁₇ H ₂₀ NSi ₂
Malic acid 3TMS	350	354	377	C ₁₃ H ₃₀ O ₅ Si ₃
	335	339	359	C ₁₂ H ₂₇ O ₅ Si ₃
	307	311	331	C ₁₁ H ₂₇ NO ₄ Si ₃
	245	249	260	C ₉ H ₁₇ O ₄ Si ₂
	233	236	251	C ₉ H ₂₁ O ₃ Si ₂
Phenylalanine 2TMS	309	-	-	C ₁₅ H ₂₇ NO ₂ Si ₂
	294	303	309	C ₁₄ H ₂₄ NO ₂ Si ₂
	266	274	281	C ₁₃ H ₂₄ NOSi ₂
	218	220	236	C ₈ H ₂₀ NO ₂ Si ₂
	192	200	201	C ₁₁ H ₁₈ NSi
Proline 2TMS	259	-	-	C ₁₁ H ₂₅ NO ₂ Si ₂
	244	249	259	C ₁₀ H ₂₂ NO ₂ Si ₂
	216	220	231	C ₉ H ₂₂ NOSi ₂
	142	146	151	C ₇ H ₁₆ NSi
Serine 3TMS	321	-	-	C ₁₂ H ₃₁ NO ₃ Si ₃
	306	309	330	C ₁₁ H ₂₈ NO ₃ Si ₃
	278	280	302	C ₁₀ H ₂₈ NO ₂ Si ₃
	218	220	236	C ₈ H ₂₀ NO ₂ Si ₂
	204	206	222	C ₈ H ₂₂ NOSi ₂
	188	190	203	C ₇ H ₁₈ NOSi ₂
Succinic acid 2TMS	262	266	280	C ₁₀ H ₂₂ O ₄ Si ₂
	247	251	262	C ₉ H ₁₉ O ₄ Si ₂

Continued on next page

Table 3.1 – *Continued from previous page*

Compound	m/z	m/z ¹³ C	m/z d ₉ -TMS	Formula
	172	176	181	C ₇ H ₁₂ O ₃ Si
Threonine 3TMS	335	-	-	C ₁₃ H ₃₃ NO ₃ Si ₃
	320	324	344	C ₁₂ H ₃₀ NO ₃ Si ₃
	218	221	236	C ₉ H ₂₄ NOSi
Tyrosine 2TMS	325	-	-	C ₁₅ H ₂₇ NO ₃ Si ₂
	310	319	325	C ₁₄ H ₂₄ NO ₃ Si ₂
	282	290	297	C ₁₃ H ₂₄ NO ₂ Si ₂
	208	216	217	C ₁₁ H ₁₈ NOSi
	192	200	198	C ₁₀ H ₁₄ NOSi
Tyrosine 3TMS	397	-	-	C ₁₈ H ₃₅ NO ₃ Si ₃
	382	391	406	C ₁₇ H ₃₂ NO ₃ Si ₃
	354	362	378	C ₁₆ H ₃₂ NO ₂ Si ₃
	280	288	298	C ₁₄ H ₂₆ NOSi ₂
	218	220	236	C ₈ H ₂₀ NO ₂ Si ₂
Uracil 2TMS	256	260	284	C ₁₀ H ₂₀ N ₂ O ₂ Si ₂
	241	245	256	C ₉ H ₁₇ N ₂ O ₂ Si ₂
Valine 2TMS	261	-	-	C ₁₁ H ₂₇ NO ₂ Si ₂
	246	251	261	C ₁₀ H ₂₄ NO ₂ Si ₂
	218	220, 222	233 236	C ₉ H ₂₄ NOSi ₂

3.2.2 TBDMS Derivatized Fragment Ions

The retained carbon atoms of all compounds in Table 3.2 can be found [Wegner et al., 2014]. Atoms that were lost in a fragment ion are shown in red and retained atoms in black.

TABLE 3.2: TBDMS Derivatized Fragment Ions

Compound	m/z	m/z ¹³ C	Formula
Alanine 2TBDMS	317	-	C ₁₅ H ₃₅ NO ₂ Si ₂
	302	305,306	C ₁₄ H ₃₂ NO ₂ Si ₂
	274	276	C ₁₃ H ₃₂ NOSi ₂
	260	263	C ₁₁ H ₂₆ NO ₂ Si ₂
	232	234	C ₁₀ H ₂₆ NOSi ₂

Continued on next page

Table 3.2 – *Continued from previous page*

Compound	m/z	m/z ¹³ C	Formula
Aspartic acid 3TBDMS	475	-	C ₂₂ H ₄₉ NO ₄ Si ₃
	460	464	C ₂₁ H ₄₆ NO ₄ Si ₃
	418	422	C ₁₈ H ₄₀ NO ₄ Si ₃
	390	393	C ₁₇ H ₄₀ NO ₃ Si ₃
	376	378	C ₁₆ H ₃₈ NO ₃ Si ₃
	316	319	C ₁₅ H ₃₄ NO ₂ Si ₂
	302	304	C ₁₄ H ₃₂ NO ₂ Si ₂
Citric acid 4TBDMS	648	-	C ₃₀ H ₆₄ O ₇ Si ₄
	633	639	C ₂₉ H ₆₁ O ₇ Si ₄
	501	507	C ₂₃ H ₄₅ O ₆ Si ₃
	459	465	C ₂₀ H ₃₉ O ₆ Si ₃
Fumaric acid 2TBDMS	344	-	C ₁₆ H ₃₂ O ₄ Si ₂
	329	333	C ₁₅ H ₂₉ O ₄ Si ₂
	287	291	C ₁₁ H ₂₀ O ₄ Si ₂
γ -Aminobutyric acid 2TBDMS	331	-	C ₁₆ H ₃₇ NO ₂ Si ₂
	316	320	C ₁₅ H ₃₄ NO ₂ Si ₂
	274	278	C ₁₂ H ₂₈ NO ₂ Si ₂
Glutamine 3TBDMS	488	-	C ₂₃ H ₅₂ N ₂ O ₃ Si ₃
	473	478	C ₂₂ H ₄₉ N ₂ O ₃ Si ₃
	431	436	C ₁₉ H ₅₃ N ₂ O ₃ Si ₃
	357	362	C ₁₇ H ₃₇ N ₂ O ₂ Si ₂
	329	333	C ₁₆ H ₃₇ N ₂ OSi ₂
Glutamic acid 3TBDMS	489	-	C ₂₃ H ₅₁ NO ₄ Si ₃
	474	479	C ₂₂ H ₄₈ NO ₄ Si ₃
	432	437	C ₁₉ H ₄₂ NO ₄ Si ₃
	358	363	C ₁₇ H ₃₆ NO ₃ Si ₂
	330	334	C ₁₆ H ₃₆ NO ₂ Si ₂
	272	276	C ₁₂ H ₂₆ NO ₂ Si ₂
Histidine 3TBDMS	497	-	C ₂₄ H ₅₁ N ₃ O ₂ Si ₃
	482	488	C ₂₃ H ₄₈ N ₃ O ₂ Si ₃
	440	446	C ₂₀ H ₄₂ N ₃ O ₂ Si ₃
	412	417	C ₁₉ H ₄₂ N ₃ OSi ₃
	280	285	C ₁₄ H ₂₈ N ₂ Si ₂
Isoleucine 2TBDMS	359	-	C ₁₈ H ₄₁ NO ₂ Si ₂
	302	304,308	C ₁₄ H ₃₂ NO ₂ Si ₂
	274	279	C ₁₃ H ₃₂ NOSi ₂

Continued on next page

Table 3.2 – *Continued from previous page*

Compound	m/z	m/z ¹³ C	Formula
	200	205	C ₁₁ H ₂₆ NSi
Leucine 2TBDMS	359	-	C ₁₈ H ₄₁ NO ₂ Si ₂
	302	304,308	C ₁₄ H ₃₂ NO ₂ Si ₂
	274	279	C ₁₃ H ₃₂ NOSi ₂
	200	205	C ₁₁ H ₂₆ NSi
Lysine 3TBDMS	488	-	C ₂₄ H ₅₆ N ₂ O ₂ Si ₃
	473	479	C ₂₃ H ₅₃ N ₂ O ₂ Si ₃
	431	437	C ₂₀ H ₄₇ N ₂ O ₂ Si ₃
	329	334	C ₁₇ H ₄₁ N ₂ Si ₂
Malic acid 3TBDMS	476	-	C ₂₂ H ₄₈ O ₅ Si ₃
	461	465	C ₂₁ H ₄₅ O ₅ Si ₃
	419	423	C ₁₈ H ₃₉ O ₅ Si ₃
	391	394	C ₁₇ H ₃₉ O ₄ Si ₃
	375	378	C ₁₇ H ₃₉ O ₃ Si ₃
	287	291	C ₁₂ H ₂₃ O ₄ Si ₂
Ornithine 3TBDMS	474	-	C ₂₃ H ₅₄ N ₂ O ₂ Si ₃
	459	464	C ₂₂ H ₅₁ N ₂ O ₂ Si ₃
	417	422	C ₁₉ H ₄₅ N ₂ O ₂ Si ₃
Serine 3TBDMS	447	-	C ₂₁ H ₄₉ NO ₃ Si ₃
	432	435	C ₂₀ H ₄₆ NO ₃ Si ₃
	404	406	C ₁₉ H ₄₆ NO ₂ Si ₃
	390	393	C ₁₇ H ₄₀ NO ₃ Si ₃
	362	364	C ₁₆ H ₄₀ NO ₂ Si ₃
	302	304	C ₁₄ H ₃₂ NO ₂ Si ₂
	288	290	C ₁₄ H ₃₄ NOSi ₂
	230	232	C ₁₀ H ₂₄ NOSi ₂
Succinic acid 2TBDMS	346	-	C ₁₆ H ₃₄ O ₄ Si ₂
	331	335	C ₁₅ H ₃₁ O ₄ Si ₂
	289	293	C ₁₂ H ₂₅ O ₄ Si ₂
	215	219	C ₁₀ H ₁₉ O ₃ Si
Tyrosine 3TBDMS	523	-	C ₂₇ H ₅₃ NO ₃ Si ₃
	508	517	C ₂₆ H ₅₀ NO ₃ Si ₃
	466	475	C ₂₃ H ₄₄ NO ₃ Si ₃
	438	346	C ₂₂ H ₄₄ NO ₂ Si ₃
	364	372	C ₂₀ H ₃₈ NOSi ₂

Continued on next page

Table 3.2 – *Continued from previous page*

Compound	m/z	m/z ¹³C	Formula
	302	304	C ₁₄ H ₃₂ NO ₂ Si ₂
Valine 2TBDMS	345	-	C ₁₇ H ₃₉ NO ₂ Si ₂
	302	304, 306	C ₁₄ H ₃₂ NO ₂ Si ₂
	288	293	C ₁₃ H ₃₀ NO ₂ Si ₂
	260	264	C ₁₂ H ₃₀ NOSi ₂

Chapter 4

Conclusion

The motivation for this thesis was to develop and apply computational mass spectrometry-based metabolomics techniques that allow to extract more biological information out of metabolomics data. First, I presented a spectrum matching algorithm that is especially suited to match compounds across different chromatograms in a non-targeted metabolomics experiment. In the context of diseases, non-targeted metabolomics methodologies have recently become more important, because cellular metabolism may be perturbed in a way that deviates from classical biochemical textbook knowledge. In this light, the ICBM algorithm can help to identify disease specific biomarkers with a higher sensitivity or can help to pinpoint targets for possible new drug treatments.

Second, I applied the ICBM algorithm to study the cellular phenotype of the human neuronal cell line LUHMES under different oxygen conditions. Although LUHMES cells should be dopamine producing and the rate limiting enzyme tyrosine hydroxylase (TH) was present in the cells, I was not able to detect dopamine in either of the two oxygen conditions. This result underlines the importance of metabolomics to study cellular phenotypes.

Third, I presented the FFC algorithm which can help to extract more biological information out of stable isotope labeling experiments. Electron ionization (EI) based mass spectrometry leads to complex mass spectra, caused by the fragmentation of the analyzed compound. The analysis of fragment ions, which contain only specific parts of the original molecule, can provide valuable information on the positional isotopic enrichment within the molecule of interest. The FFC algorithm can calculate chemical formulas and retained atoms of these mass spectral fragment ions. This information is of high interest for ^{13}C -MFA, because it provides additional constraints for the parameter fitting. Specifically, fragment ions containing different carbon atoms are of high interest, since they can carry different flux information. The FFC algorithm can complement non-targeted stable isotope-assisted methodologies. For example, the NTFD algorithm provided the

means to discover unanticipated metabolites and pathways related to specific diseases, but this information can be used only in ^{13}C -MFA if the retained carbon atoms for the fragment ion of interest are known. As such, the FFC and NTFD algorithms can help to increase the size of the metabolic network that can be profiled for ^{13}C -MFA.

Fourth, I applied the FFC algorithm to determine the chemical formulas and retained carbon atoms of 160 mass spectral fragment ions of central carbon metabolism. This fragment ion repository will facilitate the use of ^{13}C -MFA to study changes in intracellular fluxes. In particular, ^{13}C -MFA can give new insights in disease specific fluxes, as well as their regulation in central biological pathways.

Bibliography

- Antoniewicz, M. R., J. K. Kelleher, and G. Stephanopoulos (2007, January). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metabolic engineering* 9(1), 68–86.
- Chinta, S. J. and J. K. Andersen (2005, May). Dopaminergic neurons. *The international journal of biochemistry & cell biology* 37(5), 942–6.
- Fell, D. A. (2005, January). Enzymes, metabolites and fluxes. *Journal of experimental botany* 56(410), 267–72.
- Fernandez, C. A., C. Des Rosiers, S. F. Previs, F. David, and H. Brunengraber (1996, March). Correction of ¹³C mass isotopomer distributions for natural stable isotope abundance. *Journal of mass spectrometry : JMS* 31(3), 255–62.
- German, D. C. and K. F. Manaye (1993, May). Midbrain dopaminergic neurons (nuclei A8, A9, and A10): three-dimensional reconstruction in the rat. *The Journal of comparative neurology* 331(3), 297–309.
- Haverkorn van Rijsewijk, B. R. B., A. Nanchen, S. Nallet, R. J. Kleijn, and U. Sauer (2011, March). Large-scale ¹³C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. *Molecular systems biology* 7, 477.
- Hiller, K., J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, and D. Schomburg (2009, May). MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Analytical chemistry* 81(9), 3429–39.
- Hiller, K., C. Metallo, and G. Stephanopoulos (2011, July). Elucidation of cellular metabolism via metabolomics and stable-isotope assisted metabolomics. *Current pharmaceutical biotechnology* 12(7), 1075–86.
- Hiller, K., C. M. Metallo, J. K. Kelleher, and G. Stephanopoulos (2010, August). Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. *Analytical chemistry* 82(15), 6621–8.

- Hiller, K., A. Wegner, D. Weindl, T. Cordes, C. M. Metallo, J. K. Kelleher, and G. Stephanopoulos (2013, May). NTFD—a stand-alone application for the non-targeted detection of stable isotope-labeled compounds in GC/MS data. *Bioinformatics (Oxford, England)* 29(9), 1226–8.
- Holmes, E., A. W. Nicholls, J. C. Lindon, S. Ramos, M. Spraul, P. Neidig, S. C. Connor, J. Connelly, S. J. Damment, J. Haselden, and J. K. Nicholson. Development of a model for classification of toxin-induced lesions using ^1H NMR spectroscopy of urine combined with pattern recognition. *NMR in biomedicine* 11(4-5), 235–44.
- Hunter, P. (2009, January). Reading the metabolic fine print. The application of metabolomics to diagnostics, drug research and nutrition might be integral to improved health and personalized medicine. *EMBO reports* 10(1), 20–3.
- Issaq, H. J., Q. N. Van, T. J. Waybright, G. M. Muschik, and T. D. Veenstra (2009, July). Analytical and statistical approaches to metabolomics research. *Journal of separation science* 32(13), 2183–99.
- Jennings, M. E. and D. E. Matthews (2005, October). Determination of complex isotopomer patterns in isotopically labeled compounds by mass spectrometry. *Analytical chemistry* 77(19), 6435–44.
- Kim, S., I. Koo, X. Wei, and X. Zhang (2012, April). A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry. *Bioinformatics (Oxford, England)* 28(8), 1158–63.
- Kind, T., V. Tolstikov, O. Fiehn, and R. H. Weiss (2007, April). A comprehensive urinary metabolomic approach for identifying kidney cancer. *Analytical biochemistry* 363(2), 185–95.
- Koek, M. M., B. Muilwijk, M. J. van der Werf, and T. Hankemeier (2006, February). Microbial metabolomics with gas chromatography/mass spectrometry. *Analytical chemistry* 78(4), 1272–81.
- Kovats, E. (1958). Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta* 41(7), 1915–1932.
- Kusmierz, J. J. and F. P. Abramson (1994, December). Tracing ^{15}N with chemical reaction interface mass spectrometry: a demonstration using ^{15}N -labeled glutamine and asparagine substrates in cell culture. *Biological mass spectrometry* 23(12), 756–63.
- Lapainis, T., S. S. Rubakhin, and J. V. Sweedler (2009, July). Capillary electrophoresis with electrospray ionization mass spectrometric detection for single-cell metabolomics. *Analytical chemistry* 81(14), 5858–64.

- Lee, W. N., L. O. Byerley, E. A. Bergner, and J. Edmond (1991, August). Mass isotopomer analysis: theoretical and practical considerations. *Biological mass spectrometry* 20(8), 451–8.
- Lei, Z., D. V. Huhman, and L. W. Sumner (2011, July). Mass spectrometry strategies in metabolomics. *The Journal of biological chemistry* 286(29), 25435–42.
- Lorenz, M. A., C. F. Burant, and R. T. Kennedy (2011, May). Reducing time and increasing sensitivity in sample preparation for adherent mammalian cell metabolomics. *Analytical chemistry* 83(9), 3406–14.
- Macel, M., N. M. Van Dam, and J. J. B. Keurentjes (2010, July). Metabolomics: the chemistry between ecology and genetics. *Molecular ecology resources* 10(4), 583–93.
- Marin, S., W.-N. P. Lee, S. Bassilian, S. Lim, L. G. Boros, J. J. Centelles, J. M. FernAndez-Novell, J. J. Guinovart, and M. Cascante (2004, July). Dynamic profiling of the glucose metabolic network in fasted rat hepatocytes using [1,2-13C2]glucose. *The Biochemical journal* 381(Pt 1), 287–94.
- Masakapalli, S. K., P. Le Lay, J. E. Huddleston, N. L. Pollock, N. J. Kruger, and R. G. Ratcliffe (2010, February). Subcellular flux analysis of central metabolism in a heterotrophic Arabidopsis cell suspension using steady-state stable isotope labeling. *Plant physiology* 152(2), 602–19.
- McLafferty, F. W. and F. Turecek (1994, February). Interpretation of Mass Spectra, 4th ed. *Journal of Chemical Education* 71(2), A54.
- Meiser, J., D. Weindl, and K. Hiller (2013, January). Complexity of dopamine metabolism. *Cell communication and signaling : CCS* 11(1), 34.
- Metallo, C. M., J. L. Walther, and G. Stephanopoulos (2009, November). Evaluation of 13C isotopic tracers for metabolic flux analysis in mammalian cells. *Journal of biotechnology* 144(3), 167–74.
- Michelucci, A., T. Cordes, J. Ghelfi, A. Pailot, N. Reiling, O. Goldmann, T. Binz, A. Wegner, A. Tallam, A. Rausell, M. Buttini, C. L. Linster, E. Medina, R. Balling, and K. Hiller (2013, May). Immune-responsive gene 1 protein links metabolism to immunity by catalyzing itaconic acid production. *Proceedings of the National Academy of Sciences of the United States of America* 110(19), 7820–5.
- Niklas, J. and E. Heinzle (2012, January). Metabolic flux analysis in systems biology of Mammalian cells. *Advances in biochemical engineering/biotechnology* 127, 109–32.
- Niklas, J., K. Schneider, and E. Heinzle (2010, February). Metabolic flux analysis in eukaryotes. *Current opinion in biotechnology* 21(1), 63–9.

- Noguchi, Y., J. D. Young, J. O. Aleman, M. E. Hansen, J. K. Kelleher, and G. Stephanopoulos (2009, November). Effect of anaplerotic fluxes and amino acid availability on hepatic lipoapoptosis. *The Journal of biological chemistry* 284(48), 33425–36.
- Nordström, A., E. Want, T. Northen, J. Lehtiö, and G. Siuzdak (2008, January). Multiple ionization mass spectrometry strategy used to reveal the complexity of metabolomics. *Analytical chemistry* 80(2), 421–9.
- Oliver, S. G., M. K. Winson, D. B. Kell, and F. Baganz (1998, September). Systematic functional analysis of the yeast genome. *Trends in biotechnology* 16(9), 373–8.
- Patterson, A. D., H. Li, G. S. Eichler, K. W. Krausz, J. N. Weinstein, A. J. Fornace, F. J. Gonzalez, and J. R. Idle (2008, February). UPLC-ESI-TOFMS-based metabolomics and gene expression dynamics inspector self-organizing metabolomic maps as tools for understanding the cellular response to ionizing radiation. *Analytical chemistry* 80(3), 665–74.
- Raghevedran, V., A. K. Gombert, B. Christensen, P. Kötter, and J. Nielsen (2004, July). Phenotypic characterization of glucose repression mutants of *Saccharomyces cerevisiae* using experiments with ¹³C-labelled glucose. *Yeast (Chichester, England)* 21(9), 769–79.
- Sano, M., Y. Yotsui, H. Abe, and S. Sasaki (1976, February). A new technique for the detection of metabolites labelled by the isotope ¹³C using mass fragmentography. *Biomedical mass spectrometry* 3(1), 1–3.
- Schmidt, K., M. Carlsen, J. Nielsen, and J. Villadsen (1997, September). Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnology and bioengineering* 55(6), 831–40.
- Scholz, D., D. Pörtl, A. Genewsky, M. Weng, T. Waldmann, S. Schildknecht, and M. Leist (2011, December). Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. *Journal of neurochemistry* 119(5), 957–71.
- Schroeder, F. C., D. M. Gibson, A. C. L. Churchill, P. Sojikul, E. J. Wursthorn, S. B. Krasnoff, and J. Clardy (2007, January). Differential analysis of 2D NMR spectra: new natural products from a pilot-scale fungal extract library. *Angewandte Chemie (International ed. in English)* 46(6), 901–4.

- Sellick, C. A., R. Hansen, A. R. Maqsood, W. B. Dunn, G. M. Stephens, R. Goodacre, and A. J. Dickson (2009a, January). Effective quenching processes for physiologically valid metabolite profiling of suspension cultured Mammalian cells. *Analytical chemistry* 81(1), 174–83.
- Sellick, C. A., R. Hansen, A. R. Maqsood, W. B. Dunn, G. M. Stephens, R. Goodacre, and A. J. Dickson (2009b, January). Effective quenching processes for physiologically valid metabolite profiling of suspension cultured Mammalian cells. *Analytical chemistry* 81(1), 174–83.
- Spura, J., L. C. Reimer, P. Wieloch, K. Schreiber, S. Buchinger, and D. Schomburg (2009, November). A method for enzyme quenching in microbial metabolome analysis successfully applied to gram-positive and gram-negative bacteria and yeast. *Analytical biochemistry* 394(2), 192–201.
- Stein, S. E. and D. R. Scott (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Am. Soc. Mass. Spectrom.* 5, 859–66.
- Stephanopoulos, G. (1999, January). Metabolic fluxes and metabolic engineering. *Metabolic engineering* 1(1), 1–11.
- Studer, L., M. Csete, S. H. Lee, N. Kabbani, J. Walikonis, B. Wold, and R. McKay (2000, October). Enhanced proliferation, survival, and dopaminergic differentiation of CNS precursors in lowered oxygen. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 20(19), 7377–83.
- Tritsch, N. X., J. B. Ding, and B. L. Sabatini (2012, October). Dopaminergic neurons inhibit striatal output through non-canonical release of GABA. *Nature* 490(7419), 262–6.
- Vander Heiden, M. G., L. C. Cantley, and C. B. Thompson (2009, May). Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)* 324(5930), 1029–33.
- Varma, A. and B. O. Palsson (1994, October). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and environmental microbiology* 60(10), 3724–31.
- Villas-Bôas, S. G., J. Højer Pedersen, M. Akesson, J. r. Smedsgaard, and J. Nielsen (2005, October). Global metabolite analysis of yeast: evaluation of sample preparation methods. *Yeast (Chichester, England)* 22(14), 1155–69.
- Villas-Bôas, S. G., J. F. Moxley, M. Akesson, G. Stephanopoulos, and J. Nielsen (2005, June). High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *The Biochemical journal* 388(Pt 2), 669–77.

- Want, E. J., B. F. Cravatt, and G. Siuzdak (2005, November). The expanding role of mass spectrometry in metabolite profiling and characterization. *Chembiochem : a European journal of chemical biology* 6(11), 1941–51.
- Wegner, A., T. Cordes, A. Michelucci, and K. Hiller (2012). The Application of Stable Isotope Assisted Metabolomics in Biomedicine. *Current Biotechnology* 1(1), 88–97.
- Wegner, A., S. C. Săpcariu, D. Weindl, and K. Hiller (2013, April). Isotope cluster-based compound matching in gas chromatography/mass spectrometry for non-targeted metabolomics. *Analytical chemistry* 85(8), 4030–7.
- Wegner, A., D. Weindl, C. Jäger, S. C. Săpcariu, X. Dong, G. Stephanopoulos, and K. Hiller (2014, February). Fragment formula calculator (FFC): determination of chemical formulas for fragment ions in mass spectrometric data. *Analytical chemistry* 86(4), 2221–8.
- Wiechert, W. (2001a, July). ^{13}C metabolic flux analysis. *Metabolic engineering* 3(3), 195–206.
- Wiechert, W. (2001b, July). ^{13}C metabolic flux analysis. *Metabolic engineering* 3(3), 195–206.
- Wiechert, W., M. Möllney, N. Isermann, M. Wurzel, and A. A. de Graaf (1999, January). Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnology and bioengineering* 66(2), 69–85.
- Wiechert, W. and K. Nöh (2005, January). From stationary to instationary metabolic flux analysis. *Advances in biochemical engineering/biotechnology* 92, 145–72.
- Wikoff, W. R., A. T. Anfora, J. Liu, P. G. Schultz, S. A. Lesley, E. C. Peters, and G. Siuzdak (2009, March). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America* 106(10), 3698–703.
- Wishart, D. S., M. J. Lewis, J. A. Morrissey, M. D. Flegel, K. Jeroncic, Y. Xiong, D. Cheng, R. Eisner, B. Gautam, D. Tzur, S. Sawhney, F. Bamforth, R. Greiner, and L. Li (2008, August). The human cerebrospinal fluid metabolome. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences* 871(2), 164–73.
- Wu, H., A. D. Southam, A. Hines, and M. R. Viant (2008, January). High-throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Analytical biochemistry* 372(2), 204–12.

-
- Zamboni, N., S.-M. Fendt, M. Rühl, and U. Sauer (2009, January). (^{13}C) -based metabolic flux analysis. *Nature protocols* 4(6), 878–92.
- Zupke, C. and G. Stephanopoulos (1994, September). Modeling of Isotope Distributions and Intracellular Fluxes in Metabolic Networks Using Atom Mapping Matrixes. *Biotechnology Progress* 10(5), 489–498.