Aus der medizinischen Klinik mit Schwerpunkt Hämatologie, Onkologie und Tumorimmunologie und dem Molekularen Krebsforschungszentrum der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

# Analysis of metabolic fluxes in cancer cell lines with novel computational tools to enable the non-targeted tracer incorporation detection from high-resolution mass spectrometry data

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Friederike Hoffmann

aus Burg bei Magdeburg

Datum der Promotion: 30.11.2023

# Contents

## Kurzzusammenfassung

Die Möglichkeit der medizinischen Intervention im zellulären Stoffwechsel ist von großem Interesse, da sich Krebszellen diesbezüglich von normalen Zellen unterscheiden. Zudem gibt es innerhalb eines Tumors und zwischen Tumoren eine höhere genetische Heterogenität als daraus resultierende metabolische Phänotypen. Um die funktionalen Zusammenhänge vollständig zu verstehen, müssen jedoch alle Ebenen der zellulären Regulation untersucht werden, zu denen die Untersuchung der metabolischen Flüsse (Fluxomics) eine ganzheitliche, dynamische Sichtweise beiträgt. Fluxomics eröffnet neue Möglichkeiten, Krankheiten zu verstehen und neue Biomarker und Therapieziele aufzuzeigen. Massenspektrometrie und stabile Isotopenmarkierungsexperimente sind weit verbreitete Methoden, um die Flüsse kleiner Moleküle durch das metabolische Netzwerk zu verfolgen.

Hier präsentiere ich drei Software-Tools/Berechnungslösungen für einige der größten Engpässe in diesen umfassenden und datenintensiven Techniken. Das Tool *CorrectOverloadedPeaks* trägt dazu bei, die Gesamtzahl der erkannten Metaboliten pro Versuchslauf zu erhöhen, indem überladene Signale korrigiert und der dynamische Mess-Bereich erweitert wird. Metaboliten können durch Vergleich der experimentellen Daten mit Referenzspektren in Spezialbibliotheken identifiziert werden. Wenn Vergleichbarkeit und Abdeckung aufgrund technischer und biologischer Varianz gering sind, bleiben die Verbindungen "unbekannt", d. h. nicht identifiziert. Das Tool *InterpretMSSpectrum* ist in der Lage den mit verschiedenen hochauflösenden Technologien gewonnenen komplexen Massenspektren Summenformeln zuzuordnen, die auf maßgeschneiderten Regelsätzen zu chemischer Plausibilität, häufigen Addukten und neutralen Verlusten bei sanften Ionisierungstechniken beruhen. Mit *HiResTEC* steht eine empfindliche und robuste Anreicherungsberechnung für das Fluxomic-Datenhandling zur Verfügung, die wie die anderen Tools problemlos in vorhandene Datenverarbeitungs-Pipelines integriert werden kann. Es erkennt bereits Tracer-Anreicherungen von 1 % und entfernt 95 % der nicht-informativen und falsch positiven Peaks, indem ein experimentübergreifender Dekonvolutionsalgorithmus und weitere Bewertungsheuristiken genutzt werden. Es wurde umfassend mit Daten aus Krebszellkulturproben getestet und systematisch anhand vorhandener Tools und Datensätze bewertet. Es übertrifft die bestehenden Lösungen und bietet eine plattformübergreifende Kompatibilität für verschiedene hochauflösende MS-Technologien.

Alle drei Softwarepakete wurden in der Open-Source-Sprache R entwickelt und sind online frei verfügbar.


## Abstract

Targeting the metabolism is of high interest as cancer cells differ in this regard from their normal counterparts. Also, there is a higher genetic heterogeneity within a tumor and in between tumors than in resulting metabolic phenotypes. However, to fully understand the functional links, it is necessary to examine all layers of cellular regulation, to which the investigation of the metabolic fluxes (fluxomics) contributes a holistic, dynamic view. Fluxomics open up new chances in understanding diseases and thus revealing new biomarkers and therapeutic targets. Mass spectrometry and stable isotope labeling experiments are widely used methods to track the fluxes of small molecules through the metabolic network. Here I present a set of three novel computational solutions to major bottlenecks in those comprehensive and data-intensive techniques. It is shown that the tool *CorrectOverloadedPeaks* helps to increase the total number of detected metabolites per experimental run by correcting overloaded signals and extending the dynamic measuring range. Metabolites can be identified by comparison of the experimental data to reference spectra in specialized libraries. When comparability and coverage are low due to technical and biological variance the compounds remain “unknown” i.e. unidentified. The tool *InterpretMSSpectrum* is able to assign sum formulas to complex mass spectra, acquired with different high-resolution technologies, based on tailored rule sets of chemical plausibility in metabolites and common adducts and neutral losses in soft ionization techniques. With *HiResTEC* a sensitive and robust tracer enrichment calculation for fluxomics data handling is at hand, which, like the other tools, can be easily integrated into existing data handling pipelines. It detects tracer enrichment already from 1 % and removes 95 % of uninformative and false positive peaks by exploiting an experiment-wide deconvolution algorithm and further evaluation heuristics. It was rigorously tested with data from cancer cell culture samples and systematically evaluated against existing tools and data sets. It outperforms the existing solutions and provides cross-platform compatibility for different high-resolution mass spectrometry technologies. All three software packages are developed in the open source language R and are freely available online.

# Introduction

Metabolic rewiring is one of the emerging hallmarks of cancer (Hanahan and Weinberg 2011). Cancer cells, compared to normal healthy cells, have alternate routes to meet their energy demands which is known since Warburg described the effect more than 90 years ago (Otto Warburg, Karl Posener, and Erwin Negelein 1924) and renders metabolism an attractive therapeutic target. This is of high relevance since genetic alterations and acquired mutations in cancer cells are highly diverse and often unique to each tumor while resulting in a limited number of metabolic phenotypes (Martinez-Outschoorn et al. 2017). Still, most current cancer therapeutics targeting the metabolism do so on gene or protein level. But the notion that, for example, elevated mRNA or protein levels directly imply an increased function of a pathway or the influence on a cellular process is too narrowly considered and neglects the role of regulatory mechanisms and the microenvironment (Moreno-Sánchez et al. 2016). In fact, it could be shown that for a number of transcripts there is no direct correlation to the functioning enzyme or activity in the cell (Moreno-Sánchez et al. 2016; Winter and Krömer 2013), also, the protein content does not necessarily reveal to what extent the protein/enzyme is active and affecting the pathway under investigation. Furthermore, some metabolites are able to induce epigenetic changes and thereby influencing gene expression (Nowicki and Gottlieb 2015; Wishart 2016). Thus, to fully understand the cellular functional outcome, it is necessary to take all layers of regulation into account. Different from the other 'omics'-techniques metabolomics and fluxomics achieve this and open up a more holistic view on the cellular phenotype and diseases (Sauer 2006; Weindl, Cordes, et al. 2016). With $^{13}C$-based flux analysis, it is possible to quantify this integrated output of interactions and together with classical biochemistry and -analytical methods, those holistic studies can lead to a paradigm shift in how diseases are seen and thus to new options in disease diagnosis and therapy, as well as in the discovery of new targets and new drugs (Wishart 2016).

Metabolomic and fluxomic experiment generate big data sets which makes manual data evaluation no longer possible, however technical, instrumentational, and experimental set ups are diverse and existing data evaluation methods thus not necessarily cross comparable or applicable. In this study, I present a comprehensive set of computational tools to address major problems in metabolic and fluxomics experimental set-ups; namely detection limits and

dynamic range of measurements, de novo metabolite identification, and non-targeted, non-redundant, sensitive, robust, cross-platform tracer incorporation detection.

Together these tools can be integrated into any existing data handling pipeline and streamline the evaluation process and providing direct usability to flux modeling software. All are written in the open source language R www.r-project.org (R Core Team 2017) and freely available on The Comprehensive R Archive Network (CRAN, cran.r-project.org).

Alongside, to showcase the abilities of the presented computational tools that were developed in the focus of this work, I will also present the analysis of $^{13}$C-Glucose labeling experiments of the two breast cancer cell lines (MCF-7 and MDA-MB-231).

Theses cell lines are often used as models to characterize a less aggressive and transformed (MCF-7) and a more invasive and higher metastatic phenotype (MDA-MB-231). We expected that these characteristics would be detectable as differences in uptake of Glucose and Glucose metabolism and could be measured with the methods at hand and thus provide an ideal model to show the tracer incorporation over time in these cells.

MCF-7 is a well-established breast cancer cell line, derived from a pleural effusion of a metastatic mammary carcinoma from a 69-year-old Caucasian, female cancer patient in 1970 (COMŞA, CÎMPEAN, and RAICA 2015; DSMZ n.d.)

It is characterized as a poorly aggressive and non-invasive cell line (Shirazi et al. 2011) with a low metabolic potential (Gest et al. 2013). Despite some heterogeneity in this cell line and clonal variants, the cells are overall considered Estrogen receptor (ER) and progesterone receptor (PR) positive and express epidermal growth factor receptor (EGFR) and the human epidermal growth factor receptor-2 (HER2) amplification, all of which are associated with a favorable clinical outcome of breast cancer as they are responsive to hormone treatment.

In mice MCF-7 cells do not induce metastasis and show a low migratory and invasive phenotype. Together with a low angiogenic potential this cell line is described as lacking tumorigenicity (Aonuma et al. 1999)

MDA-MB-231 cells were derived from a metastatic site of a breast adenocarcinoma of a 51-year-old Caucasian female in 1973 (DSMZ n.d.). It is a highly invasive, aggressive and poorly differentiated cell line. As a triple negative breast cancer cell line, it lacks the expression of the afore mentioned receptors ER, PR and HER2 amplification. (Chavez, Garimella, and Lipkowitz 2010; European Collection of Authenticated Cell Cultures 2017). Triple negative breast cancers are associated with a worse prognostics and limited therapeutic options, late and early stages

are treated commonly with chemotherapy as a receptor targeted therapy is not possible. MDA-MB-231 is often used for a late state cancer model. In mice it forms spontaneous metastatic sides in lymph nodes (Welsh 2013).
This study shows the qualitative differences in $^{13}$C-Glucose metabolism between MCF-7 and MDA-MB-231 by using the newly developed data evaluation pipeline and describes the potential and importance to further advance the field of metablomics and fluxomics.

Metabolites are a very heterogenic group of compounds regarding physicochemical properties and concentration range in biological samples (ranging from picomolar to millimolar as annotated in the Human Metabolome Database (HMDB) (Wishart et al. 2013)). Mass spectrometry (MS) coupled separation methods are ideal technologic platforms to cope with this diversity. MS coupled to high-performance chromatographic separation systems (usually Liquid Chromatography (LC) or Gas Chromatography (GC)) is one of the most sensitive and selective tools available and is broadly applicable to many compound classes (Dunn 2008; Dunn et al. 2013; Strehmel et al. 2014). Atmospheric pressure chemical ionization (APCI) has been introduced more than 40 years ago (Horning et al. 1973, 1977) but only recently has found its way into routine use in Metabolomics. Availability and technological progress made APCI-MS one of the emerging analytical systems. Superior sensitivity, detection limits, dynamic ranges, and speed (Carrasco-Pancorbo et al. 2009; Dunn et al. 2013; Dunn, Bailey, and Johnson 2005; Wachsmuth et al. 2011, 2015) are the main improvements over other established mass spectrometric set-ups like electron impact (EI) and are of special interest in non-targeted metabolomics assays.
Quantifying all the detected signals in a sample however remains challenging even with modern highly sensitive MS instruments. While on the one hand, the increased sensitivity of modern MS instruments enables the detection of low abundant molecules, and thereby the detection of possible new (bio-) markers, on the other hand, it leads to saturation of the mass detector for high abundant compounds. Classically, this can be resolved with dilution series of the sample and/or several measurements. In praxis, however, sample material is often sparse and analytical time costly, which renders an additional experiment inefficient or even impossible.
It is shown that a computational approach can extend the dynamic range of GC-APCI-measurements on average by one order of magnitude. This enables the detection and analysis of more metabolites in one experimental run.

In non-targeted metabolomics, especially, metabolite identification remains one of the major bottlenecks. When experimentally obtained spectra cannot be annotated with the help of libraries due to differences to or lack of reference, compounds remain unidentified or only roughly classified according to their chemical properties (Tsugawa et al. 2011). APCI as soft ionization technique opens new opportunities in *de-novo* annotation of "unknowns". It offers advantages for the analysis of labile compounds or compounds difficult to ionize, as, compared to EI, no intense fragmentation takes place. Most importantly, the information of the protonated molecules ($[M + H]^+$) is preserved and can be used for identification and sum formula elucidation (Jaeger et al. 2016). In Jaeger et al. 2016 a software tool that automates precursor and fragment detection with a GC-APCI tailored rule set is presented by relying on common neutral losses or adducts it assigns ranked plausible sum formulas, compares to metabolic databases, where possible, and generates informative graphical output.

Fluxes cannot be measured directly but need to be calculated from changes in metabolite levels. Thus, it is necessary to quantify the conversion of metabolites in the network. Stable isotope labeled substrates are commonly used as tracers for these analyses.

The $^{13}C$ isotope is frequently used as a tracer since all biological compounds contain carbon in a significant amount; further $^{15}N$ is commonly used to study the nitrogen metabolism. The tracer incorporation is monitored through changes in the mass isotopomer distribution (MID), as the heavier $^{13}C$ Carbon isotope accumulates in the metabolite pool. The MID, also called mass distribution vector (MDV), describes the relative intensity of all measured isotopologues per metabolite [Figure 1]. Mass isotopomers or isotopologues are defined as compounds that only differ in their isotopic composition

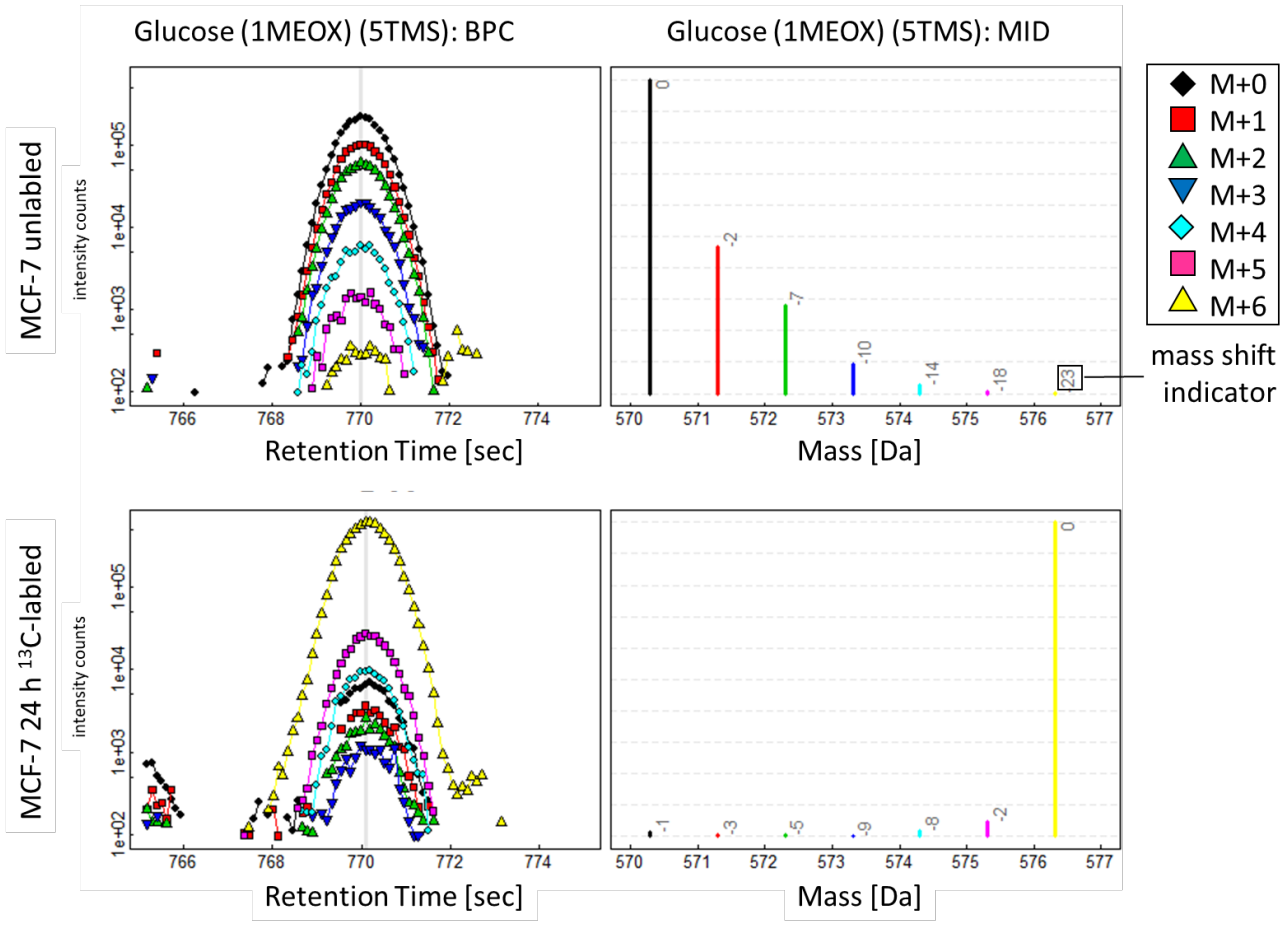



*Figure 1 Example plot of a base peak chromatogram and a MID of Glucose peak before and after 13C-labeling for 24 hours in the human breast cancer cell line MCF-7. The naturally occurring isotopes form a flute-like distribution (M+0 to M+6), after 24 hours of labeling almost all un-labeled Glucose is metabolized and only the fully labeled molecules (M+6) remain.*

All compounds containing the tracer need to be detected in the experimental data set, to allow enrichment calculation in the following. Enrichments of the tracer in the compounds are calculated by the amount of $^{13}C$ divided by total C in the MID (Fisher, Haines, and Volk 1979). The obtained enrichment data then can be mapped to a detailed mathematical model of the metabolic pathways accounting for the stoichiometry of the reactions, mass isotope distributions and carbon atom transitions (Sauer 2006; Wiechert 2001; Wiechert et al. 2001; Zamboni 2011; Zamboni et al. 2009). In classic flux balance analysis, a so-called solutions space for the metabolic fluxes (Buescher et al. 2015) is calculated. In $^{13}C$-fluxomics the models can be further constrained and algorithms start with arbitrary assumed flux values and fit the experimental data and the flux values to the least residual error (Winter and Krömer 2013). Derivatization is necessary before analyzing compounds with gas chromatography that are thermic unstable or not volatile. Here, polar groups are substituted by less polar groups to render the compounds more volatile and less labile. Often methoximating and silylating reagents are used, those introduce significant amounts of Silicon and Carbon atoms of non-

biological origin to the analyte compound. At a resolution of about R~ 50.000 which is typical for current TOF instruments the fine structure of isotopologues containing different mixtures of Si and C isotopes with a similar nominal mass cannot be resolved. Instead, a mass shift occurs away from the expected mass of the isotopologue in question, which can render the detection of the correct peaks difficult when not allowing for the right mass window. Figure 1 shows an example, annotating those detected mass shifts as small numbers in the MID. The effects of mass shift and drift will be discussed in more detail in (Hoffmann et al. 2018) and the corresponding supplemental material. Software solutions for tracer enrichment detection and calculation have to take these specific features in GC-MS into account. As the published tools at time did not meet those requirements, especially for high-resolution GC-APCI-data, a solution in the open-source scripting language R was developed. The tool *HiResTEC* (high-resolution Tracer Enrichment Calculation) addresses major points in un-targeted GC-APCI fluxomic data evaluation: sensitive tracer incorporation detection (< 1%), GC and LC compatibility, non-redundant candidate list by spectral correlation and providing a graphical output for quick and easy visual quality control.

# Methods

## Cell culture Methods and Sample Preparation

Cancer cell lines MDA-MB-231 was obtained from Charité - Universitätsmedizin Berlin Institute of Pathology, Lab for Molecular Tumor Pathology, MCF-7 from DSMZ, and grew under standard conditions: DMEM (GIBCO) supplemented with 10 % FBS (Sigma) and 1 % Penicillin/Streptomycin (Corning) in a humidified incubator at 37 °C and 5 % $CO_2$.
For $^{13}C$-labeling experiments, 24 hours before harvest/quenching cells were seeded in 6-well plates á $0{,}25x10^6$ cells per well and medium was changed to DMEM with 4,5 g/L U-$^{13}C$-Glucose (Sigma) according to the planed labeling duration (5 min, 15 min, 24 h, for the breast cancer examples). Prior to harvest cells were washed twice with 0,9 % NaCl and quenched and fixed with -80 °C Methanol (Biosolve Chemicals). Cells were scraped off, re-suspended and transferred to micro reaction vials, cell debris was pelleted by centrifugation, and aliquots of the supernatant were transferred to conical glass vials and vacuum dried in a freeze-dryer (Christ).
Lymphoma cell culture samples were received as cell pellets and extracted as described above.

## GC-MS-Processing

Dried methanolic extracts were derivatized online using 10 µl Methoxyamine (20 mg/mL in pyridine; Sigma), and 20 µl N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA, Macherey-Nagel) for 90 and 30 minutes, respectively, at 34 °C before injection of 1 µl with a split ratio of 10 % by an RTC PAL System. Data was recorded at a scan rate of 10 Hz using a Bruker Impact II mass spectrometer (resolution: ~35,000). Detailed acquisition parameters can be found in SI Text Table S2 of (Hoffmann et al. 2018).

## Data Processing

Raw data files from the MS measurements were exported to mzXML file format and further processed as described in detail in the enclosed publications (Hoffmann et al. 2018).
In short, prior to peak picking, grouping and retention time alignment which was performed using the R package *xcms* (Smith et al. 2006) package sample files were corrected for overloaded peaks with *CorrectOverloadedPeaks* (Lisec et al. 2016).Parameter settings were as follows for xcms for evaluation of GC-APCI example data set: method="centWave", ppm=25, peakwidth=c(1,6), snthresh=1, prefilter=c(5,2000) and noise=1 for function *xcmsSet* and

minsamp=6, bw=0.5 and mzwid=0.25 for function *group*. And for *CorrectOverloadedPeaks* method= *"Isoratio"*.

The functionalities of the processing steps with in the *HiResTEC* package are described in detail in the corresponding publication and will thus be only highlighted briefly in the following.

In *EvaluatePairsFromXCMSset* the from previous steps resulting *xcmsSet* object is scanned for peaks that differ in multiples of the mass difference of $^{13}C$ and $^{12}C$ Carbon (n*1.003355 Da, for Carbon labeling experiments) on the mass scale. *EvaluateCandidateListAgainstRaw* encompasses several functions, summarized in the following. *RankCandidateList*, which sorts the list of matching (mass-charge) m/z-pairs descending by the sum of their intensity over time, and thereby enable the evaluation of major peaks first.

*EvaluateCandidate* extracts base peak chromatograms (BPCs) experiment wide for the peak and determines the enrichment and the enrichment change over the experimental time and using an ANOVA model to test the statistical significance. Along with other quality checks *DeconvoluteSpectrum* and *EvaluateSpectrum* detect spectral correlation over all samples and within a sample, thereby enabling the detection of peak fragments that have already been evaluated and thus, do not need further attention. The remaining candidates are summarized with the calculated enrichment and statistical values in an Excel spreadsheet and for visual monitoring the package provides a pdf-document, containing spectral information, BPCs, box and scatter plots on the enrichment information. The output list can be used to identify candidates and, if an unambiguous sum formula can be assigned or is available from a target library, for the correction of the MID for natural occurring isotopes, which makes the data directly usable for flux modelling attempts.

## Results

In this study I present three software packages, freely available on CRAN repository, providing computational solutions for data handling, statistical evaluation and interpretation for metabolomics and fluxomics workflows. Together addressing major bottle necks in these kinds of experiments and providing a coherent, robust and sensitive data evaluation pipeline, especially suited for GC-APCI-MS based non-targeted fluxomics.

*CorrectOverloadedPeaks* uses two different approaches to estimate the peak intensity of signals reaching the upper limit of detection: Gauss curve approximation (G) and isotopic ratio (IR).

The two systems perform differently and have different advantages for different applications. First, for signals approaching the detector saturation the software extracts all BPCs in narrow retention time frames. Then the overloaded data points are removed and corrected by the algorithm.

Gauss approximation fits a Gauss curve based on the front and back of the peak signal with the least residual error. It could be shown that those data points maintain the geometric properties of the curve and thereby allow mathematical fitting. IR uses the first isotope not reaching the saturation and calculates the ratios of the isotopic traces in the front and tail of the peak and corrects the missing apex by using those stable ratios.

The linear range (LR) was determined for the measurement of a standard mix of 62 metabolites. All LRs were compared and statistically evaluated, before and after computational correction. Though IR results in a lower median LR gain of 0.6 orders of magnitude, where Gauss gains 1.4 orders of magnitude; IR handles skewed peak shape, fronting and tailing more stable than Gauss curve approximation, if enough data points in the front or back are available. Furthermore, preserving the precise isotopic ratio is crucial for tracer incorporation calculation in fluxomics experiments, thus for those data sets the IR methods has to be used. For two thirds of metabolites in the test mixture more than 50 % of the potential LR gain could be reported, independent of substance class. Plots of all analyzed and corrected metabolite peaks and data to the specific linear ranges and gains can be found online in the supplemental material of (Lisec et al. 2016).

The peak correction resulted in low residual errors (< 20%) in over 90 % of the analyzed peaks, both in the dilution series of a metabolite mix and in an analysis of metabolites in a biological matrix, here blood serum. Using *CorrectOverloadedPeaks* the total number of detected metabolites can be time- and cost efficiently increased, without the need of additional (wet lab) experiments.

The information of the molecular ion, frequently preserved in GC-APCI-MS, together with specific in-source fragments and typical adducts or neural losses, can be used to annotate the mass spectra and assign sum formulas to measured compounds.

According to the mass of the peak and the chemical elements occurring in biological compounds, for each peak a set of possible chemical element combinations can be calculated. Not all of those mathematically correct combinations are meaningful in a biological or chemical sense. To reduce the list to more plausible suggestions, a rule set for typical elemental ratios

and combinations was derived from the entries in the Golm Metabolome Database (GMD) (Kopka et al. 2005). Further, a set of common losses specific for GC-APCI-MS was added to the filter rules. Filtering the primary suggestions by the elemental composition reduced the list by 89 % on average. The most common neutral losses were $CH_4$, TMS-OH and O-DMS, the implementation of this information reduced the list further by additional 98 % on average. Those rule sets can be modified according to the user's requirements.

Working on a standard mixture of 59 metabolites, *InterpretMSSpectrum* ranks correct sum formula on place one, for 84 % and on 2-3 in further 7 % of the cases. The full set of annotated compound spectra can be found online in the supporting information of (Jaeger et al. 2016).

The comparison of different deconvolution tools, for data preprocessing, showed significant performance differences in spectral annotation, when using *InterpretMSSpectrum*. Here, the algorithm showed to be also a valuable quality check tool for data pre-processing.

The objective for developing *HiResTEC* was the fast, sensitive, robust, and nonredundant detection of tracer enrichment in non-targeted fluxomics.

The package provides potent filter heuristics, described in the following, resulting in a reliable enrichment detection down to 1 % $^{13}C$ and removes over 95 % of false positive hits and redundant information from the candidates list.

The data of 36 Lymphoma cell culture samples was exemplarily analyzed and is presented in the publication.

Additionally, I conducted $^{13}C$ -Glucose labeling experiments on two breast cancer cell lines, MCF-7 and MDA-MB-231, to characterize their metabolic (flux) differences and to illustrate the main steps of the data evaluation flow of the algorithms. A selection of this data of MCF-7 and MDA-MB-231 cell culture samples in triplicates during three labeling time points (5 min, 15 min, 24 h) is shown in comparison below, analyzed step by step and discussed.

Though many labeling experiments were conducted also with the aforementioned cell lines coverage and data quality to draw biological relevant conclusions was deemed insufficient thus those will not appear in the presented work. These data points are not discussed in detail, but contributed to the development, establishment and testing of the presented packages, especially determination of data quality, data quality check filters and pattern, determination of error rates and optimization of plotting layouts.

Regardless of the small size of the remaining data set, which does not allow for in depth correlation and holistic fluxomic analysis, distinct patterns and trends of enrichment could be

found between the two cell lines, which support the overall strength and importance of metabolomic and fluxomic research in the medical field.

The fundamental data pre-processing is described in the *Methods* section. The generic data files from the MS manufacturer (Bruker Daltonics) were converted to mzXML. In a table information on the sample ID, cell line, labeling duration (TP), replicate, raw data file path etc. is given. With this information a list of all raw data files is loaded into the working memory.
Working on those raw data files, first, a plot of all chromatograms is generated before the data is subjected to any other processing steps. The output of the function *VisualChromatogramInspection* provides an overview on all files at once, and their quality and possible problems and gross differences can be assessed. Figure 2 shows the overview for the 24 selected samples.

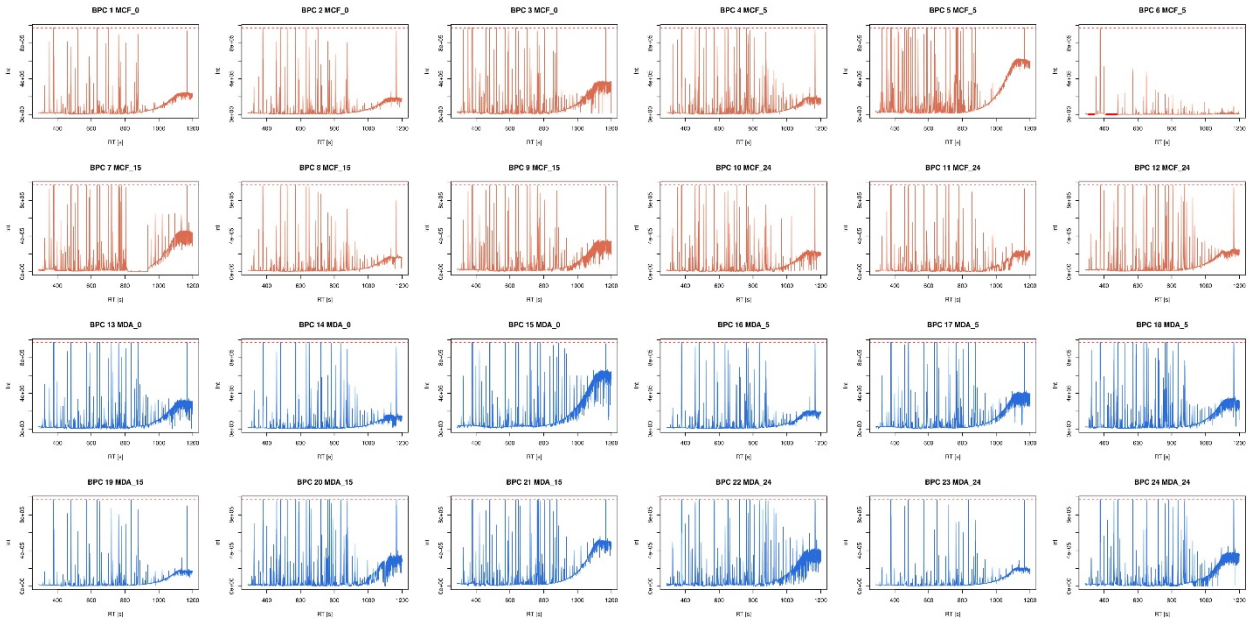



*Figure 2 Overview plot of BPCs of all samples in the set before overloading correction. Red plots are MCF-7 cell samples. Blue plots are MDA-MB-231 cell samples. Y-axis intensity. X-axis: RT (260-1260). The red dashed line marks the detector saturation (ds = 971775).*

In this example it is apparent that sample number 6 has less peaks and less intensity than the other files, and should be omitted from further evaluation, as it could add variance or unreliable data points to the peak set under evaluation.
Next, the peaks in the data reaching the upper limit of detection are going to be corrected using *CorrectOverloadedPeaks*. After this the BPCs of the samples can be checked again in an overview plot, as in Figure 3.

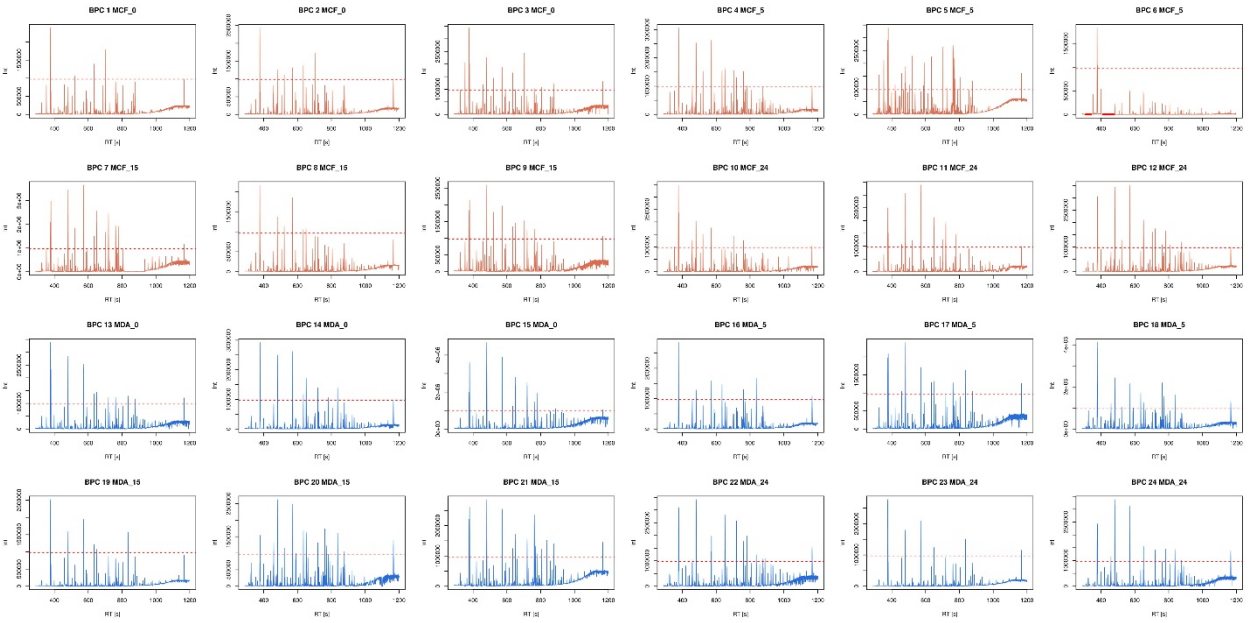



*Figure 3 Overview plot of BPCs of all samples in the set after overloading correction. Red plots are MCF-7 cell samples. Blue plots are MDA-MB-231 cell samples. Y-axis: intensity. X-axis: RT (260-1260). The red dashed line marks the detector saturation (ds = 971775).*

*CorrectOverloadedPeaks* provides a pdf file containing plots of all corrected peaks and saves a raw data file with the corrected peak intensity values. Figure 4 shows three examples of the result of the correction in three differently strong overloaded peaks.

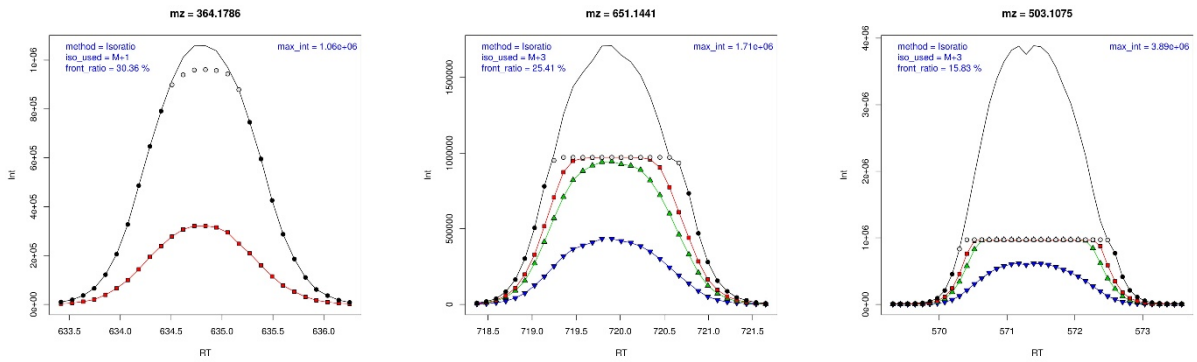



*Figure 4 Plots of peaks that were corrected by the algorithm CorrectOverloadedPeaks in sample MDA-MB-231, TP=0, replicate 3 The gray dots mark the original data, while the black lines represent the result of correction. Different extents of overloading can be corrected, depending on that a heavier isotopologue has to be used which did not reached the detector saturation. Which iosotopologue was used for the correction is given in the top left corner of the plot along with the used method. In the top right corner, the corrected maximal intensity is given.*

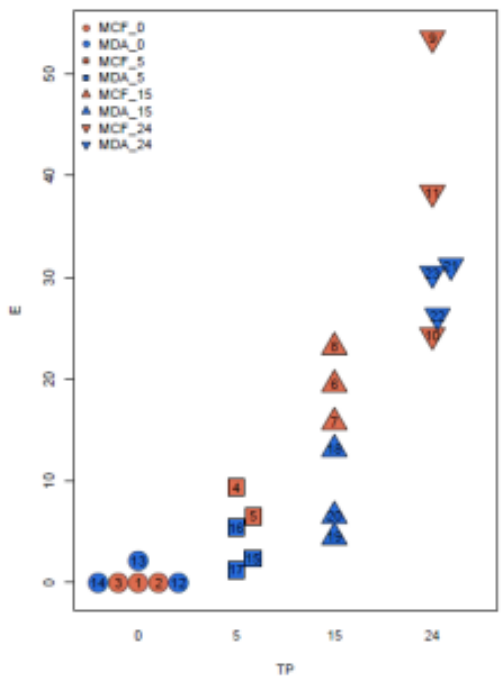
After this the data can be evaluated targeted, using mass and RT information of known compounds, or non-targeted, without using a compound library. The targeted evaluation is faster, as only specific peaks must be processed by the algorithm. Also, the identity and sum formula are necessary information for the later MID correction and thus flux modelling attempts. The non-targeted search, however, is explorative and might result in the detection of

compounds, not present in target lists, that show significant patterns in tracer incorporation and the metabolic network. Those could be possible new biomarkers or therapy intervention points.

The following intermediate steps of the targeted search do not generate graphical output until the end. The script extracts BPCs, plots them, determines the base formula from the information in the target compound library, extracts MIDs, corrects for natural occurring isotopes, calculates the tracer enrichment from those and last, generates a set of plots from this data for each metabolite and saves it as a pdf file. In Figure 5 the scatter plots of the tracer enrichment of each sample for some metabolite plots are shown. A tracer enrichment over time can be observed in both cell lines, but to a different extent and velocity.

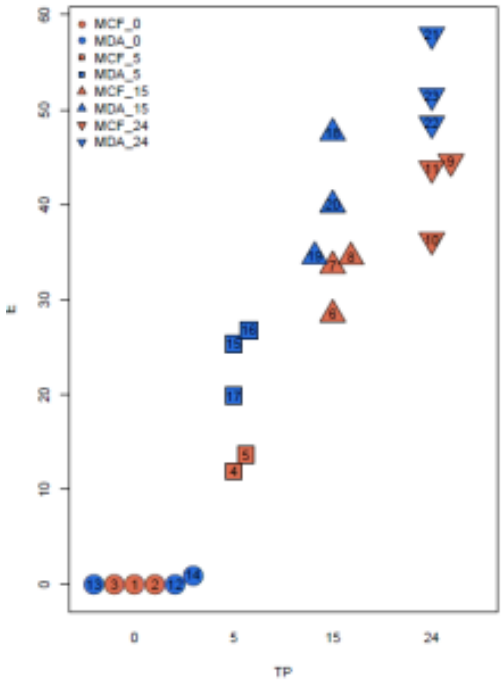
A

Pyruvic acid




B

Lactic acid



C
Malic acid

D
Citric acid

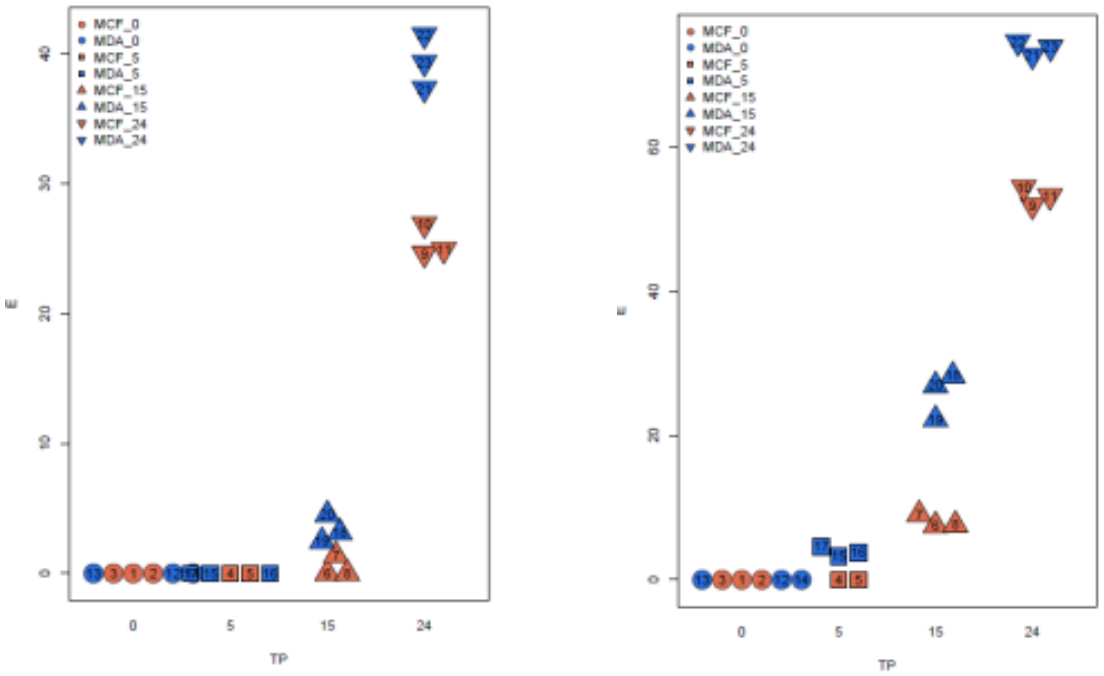



*Figure 5 Scatter plots of selected metabolites from the targeted search. Red for MCF-7 samples. Blue for MDA-MB-231 samples, and symbols according to labeling duration. A Pyruvic acid (-CH4) (C3H2NO3). B Lactic acid (2 TMS) (C3H5O3). C Malic acid (3 TMS) (C4H5O5). D Citric acid (4 TMS) (C6H7O7)*

The Peak shape and data quality in the QC Plots of Asparagine (4TMS) BP1 and Uracil (2TMS) has led to the exclusion of these two compounds from further analyses.

The complete data and print outs exceed the limit of the printed study and can be found in the electronic version.

The output of the evaluation script provides a list of relative enrichments of each target compound. Additionally, the velocity (or kinetics) of the enrichment was calculated dividing the relative enrichment by the labeling time. The complete list including calculations of the following analyses can be found in the electronic version.

Figures Figure 6 to Figure 9 show the median enrichment over time and the median enrichment velocities over time for both cell lines, respectively. While the speed of enrichment is rapidly decreasing after 15min and diverges towards a steady state (metabolic steady state), the level of enrichment reaches a saturation towards 24 h of labeling time, in both cell lines, reversely proportional to the velocity of enrichment, as expected.

In the targeted enrichment evaluation metabolites from glycolysis, citric acid cycle, amino acid metabolism, purine and pyrimidine metabolism and fatty acid metabolism are found enriched with $^{13}C$ carbon. Enriched compounds belong to the substance classes sugar alcohols, organic acids, amino acids, phosphor esters, and nucleo bases, in short, frequently found metabolism intermediates stemming from glucose break down through glycolysis and adjacent metabolic pathways. The majority of enriched compounds is belonging to organic acids, sugar alcohols and amino acids [Table 1].

Through methionine salvage pathway even Methionine could be found enriched (62.7 % in MCF-7 and 45.0% in MDS-MB-231), though being an essential amino acid in mammals.

Many sugar derivates can also be found enriched, despite not being classical members of Glycolysis or TCA cycle and similar. Molecular rearrangements in liquid environments and in the following chemical substitution with TMS groups the molecular structure is fixed and thus identified as such.

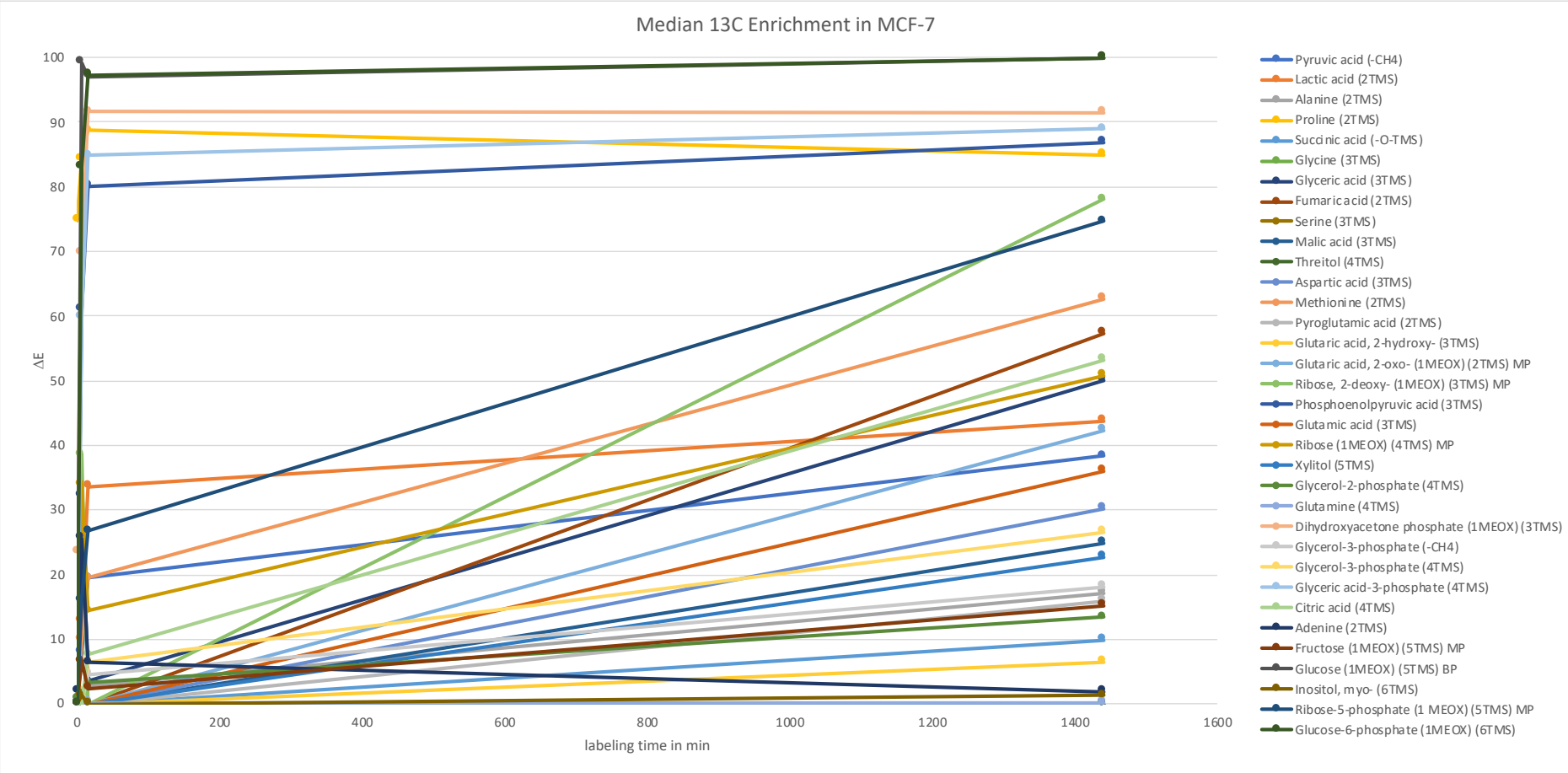



*Figure 6 Relative median $^{13}C$ enrichment over labeling time in minutes of all metabolites in the targeted analysis of MCF-7*

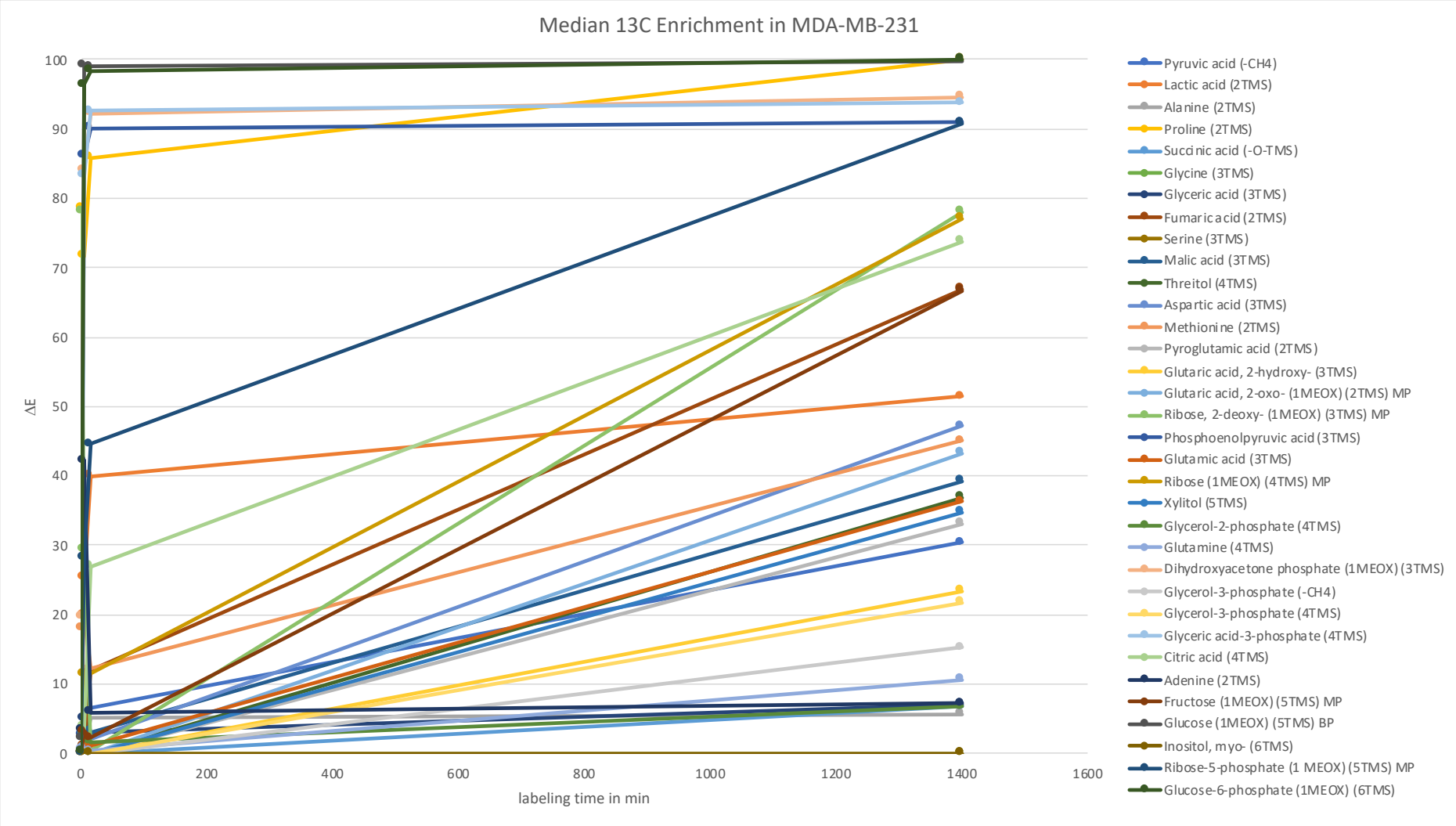



*Figure 7 Relative median $^{13}C$ enrichment over labeling time in of all metabolites in the targeted analysis of MDA-MB-231*

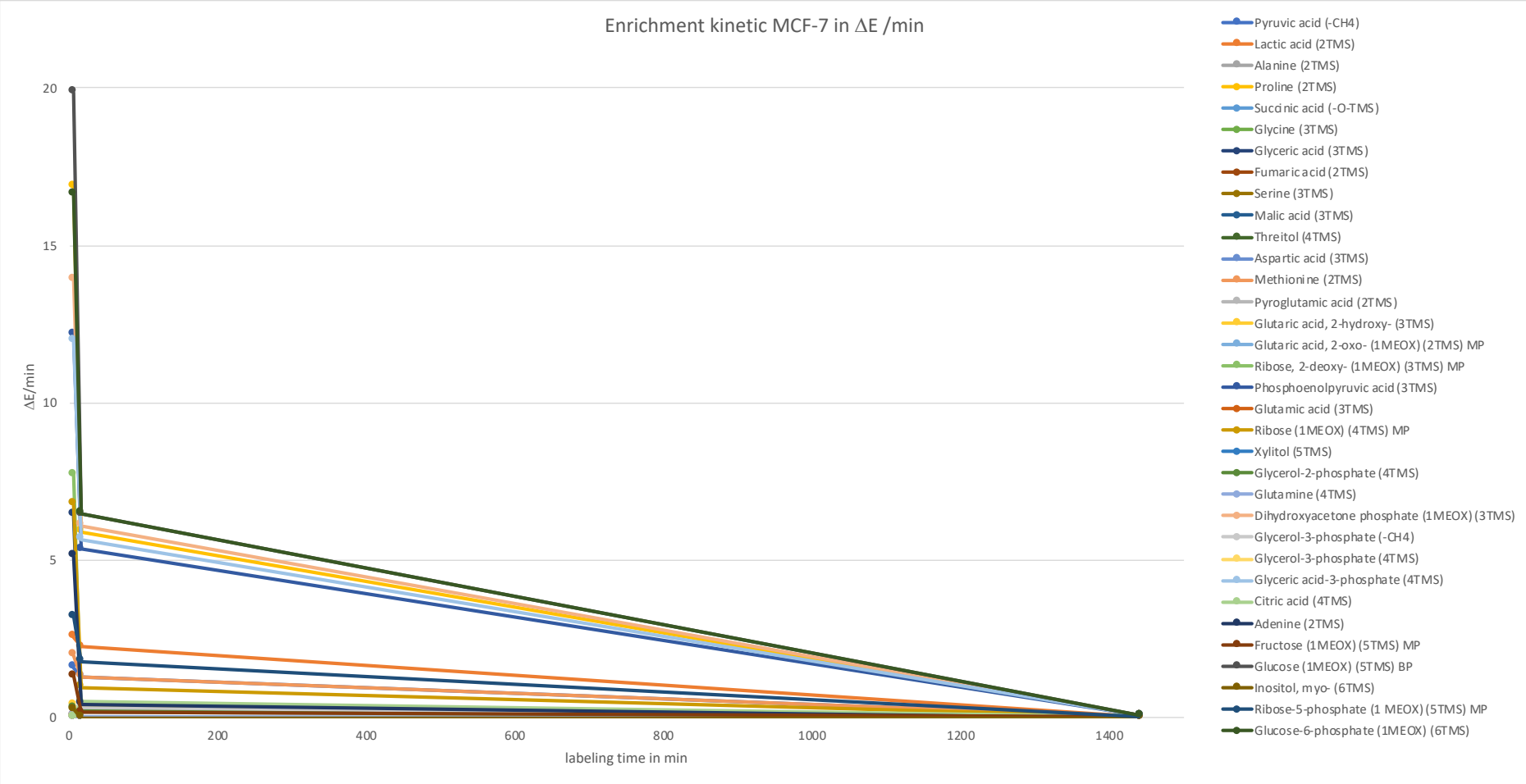



*Figure 8 $^{13}C$ enrichment kinetics in delta E/min over labeling time for all metabolites in the targeted analysis of MCF-7*

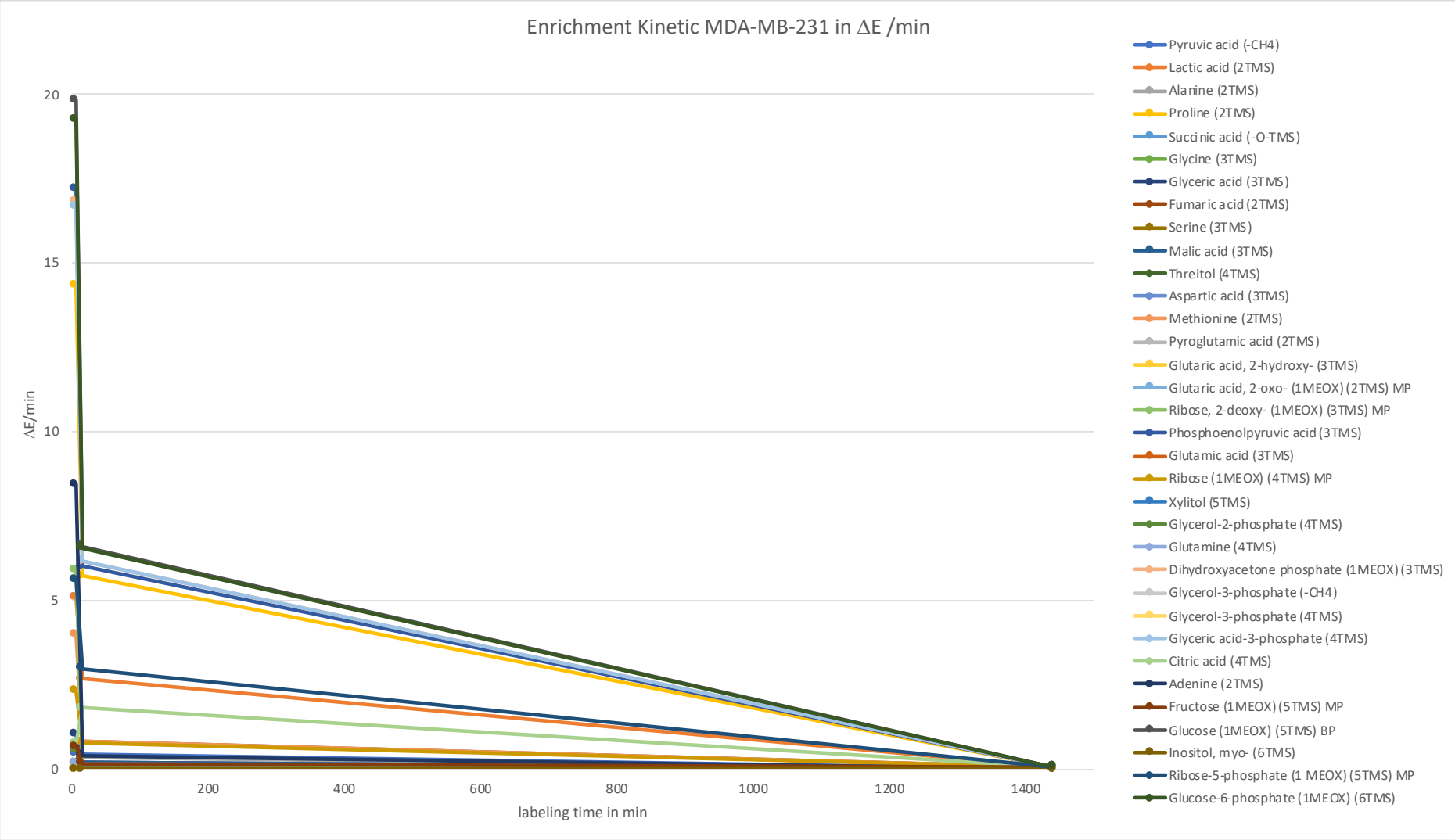



*Figure 9 $^{13}C$ enrichment kinetics in delta E/min over labeling time for all metabolites in the targeted analysis of MDA-MB-231*

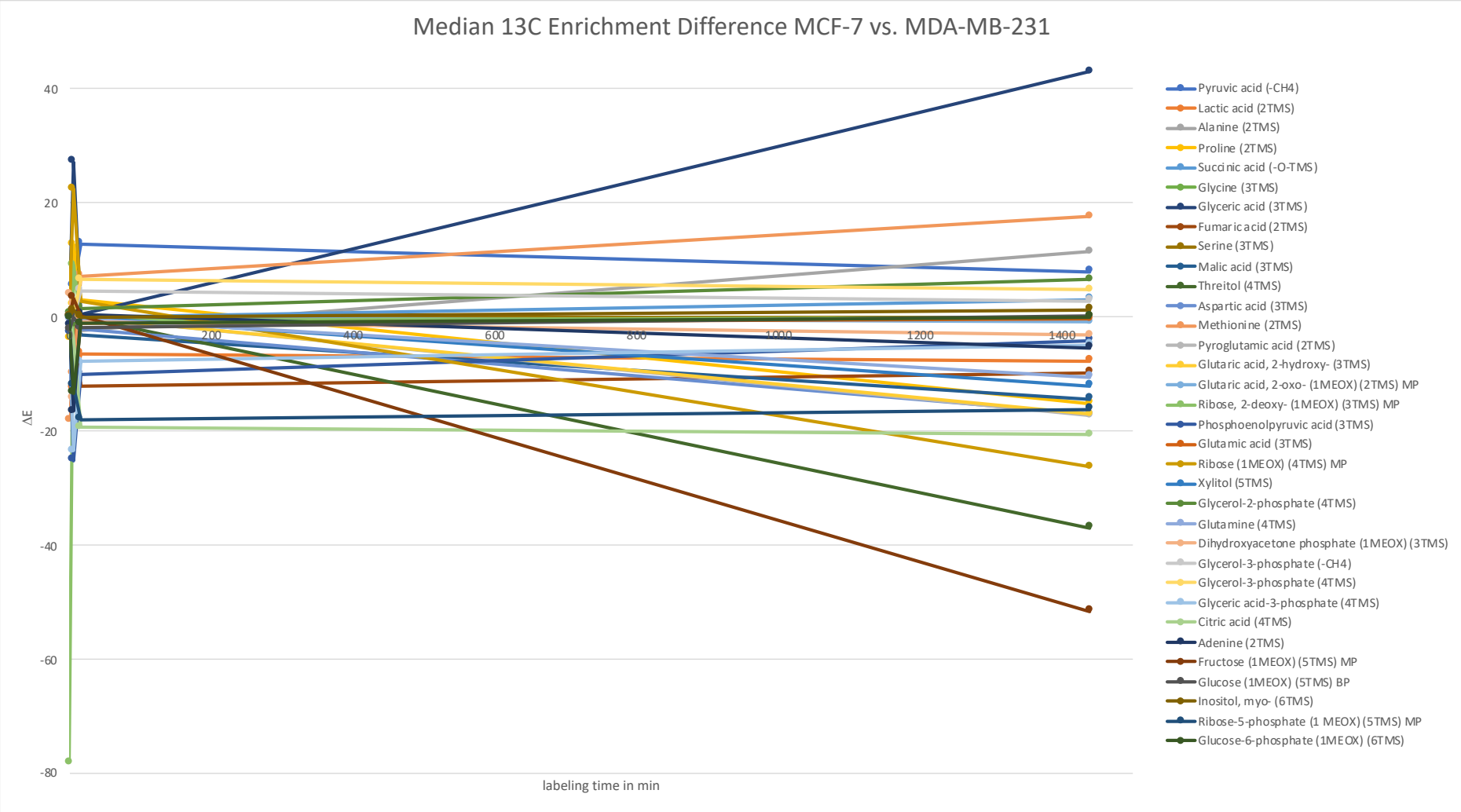



*Figure 10 Difference in median $^{13}C$ enrichment between MCF-7 and MDA-MB-231 over labeling time in min. Difference dE = median dE(MCF-7) – median dE(MDA-MB-231)*

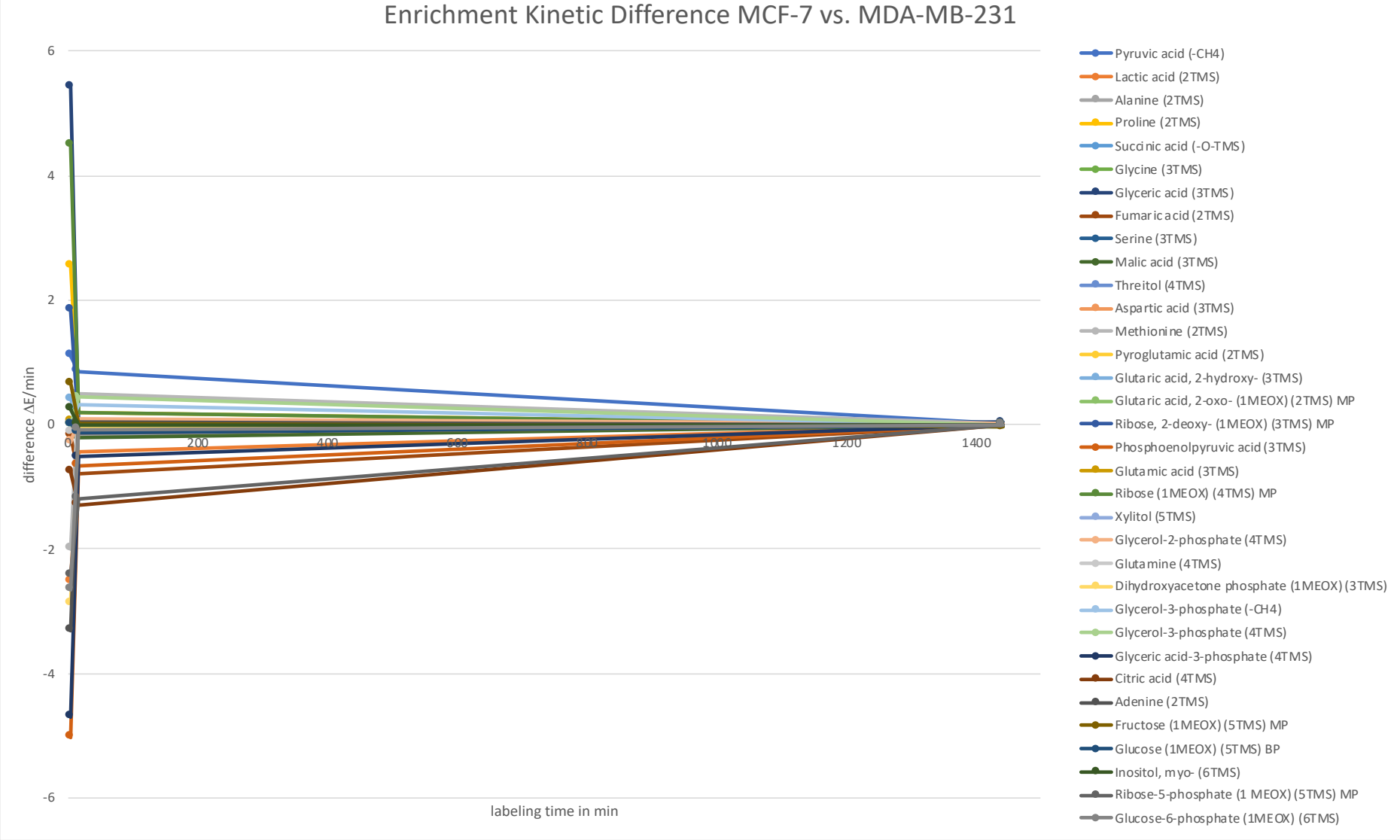



*Figure 11 Difference of median $^{13}C$ enrichment kinetcis between MCF-7 and MDA-MB-231 over labeling time in min. Difference dE/min = median dE/min (MCF-7) – median dE/min (MDA-MB-231)*

For an easy evaluation of the qualitative differences in enrichment and enrichment velocity the difference between the enrichment values and enrichment velocity values was calculated and plotted in Figures Figure 10 and Figure 11. Comparison of those between the cell lines and time points resulted in the following list of metabolites with major differences in $\Delta$E (absolute difference $\Delta$E > 10%) Table 1.

*Table 1 Summary of direct comparison of enriched compounds between MCF-7 and MDA-MB-231*

| Compounds with at least 10% higher enrichment in MCF-7 after 24h labeling | Compounds with at least 10% higher enrichment in MDA-MB-231 after 24h labeling |
|---|---|
| Glyceric acid (3TMS) | Citric acid (4TMS) |
| Methionine (2TMS) | Ribose (1MEOX) (4TMS) MP |
| Alanine (2TMS) | Ribose-5-phosphate (1 MEOX) (5TMS) MP |
| | Fructose (1MEOX) (5TMS) MP |
| | Proline (2TMS) |
| | Fumaric acid (2TMS) |
| | Malic acid (3TMS) |
| | Threitol (4TMS) |
| | Aspartic acid (3TMS) |
| | Pyroglutamic acid (2TMS) |
| | Glutaric acid, 2-hydroxy- (3TMS) |
| | Xylitol (5TMS) |
| | Glutamine (4TMS) |

For comparing the velocity values of both cell lines, a cut-off of an absolute difference in $\Delta$E/min > 1% was chosen to determine the list of metabolites with most noticeable different enrichment velocity. Results are summarized in Table 2.

*Table 2 Comparison of enrichment velocities between MCF-7 and MDA-MB-231*

| Compounds with higher enrichment velocity ($\Delta$E/min) in MCF-7 after 5 min labeling | Compounds with higher enrichment velocity ($\Delta$E/min) in MDA-MB-231 after 5 min labeling |
|---|---|
| Pyruvic acid (-CH4) | Lactic acid (2TMS) |
| Proline (2TMS) | Methionine (2TMS) |
| Glyceric acid (3TMS) | Phosphoenolpyruvic acid (3TMS) |
| Ribose, 2-deoxy- (1MEOX) (3TMS) MP | Dihydroxyacetone phosphate (1MEOX) (3TMS) |
| Ribose (1MEOX) (4TMS) MP | Glyceric acid-3-phosphate (4TMS) |
| | Adenine (2TMS) |
| | Ribose-5-phosphate (1 MEOX) (5TMS) MP |
| | Glucose-6-phosphate (1MEOX) (6TMS) |

MCF-7 has in direct comparison fewer metabolites higher enriched than MDA-MB-231. MCF-7 exceeds the enrichment level of metabolites only in three cases (Glyceric acid, Methionine, and Alanine) while MDA-MB-231 has 13 metabolites higher enriched than MCF-7 (Table 1).
In case of enrichment speed MDA-MB-231 shows faster velocities in more metabolites (8) compared to MCF-7 (5). The metabolites where one cell line is faster in enrichment are not necessarily the metabolites which this cell line enriches to a higher level.
In substance classes there is no distinct trends detectable which class would be enriched more by one of the cell lines or in general. To come to a reliable conclusion the sample size is too small in differentially enriched metabolites and differences in velocities.
Overall, it can be determined that MDA-MB-231 metabolizes faster and enriches metabolites seemingly higher than MCF-7. The afore mentioned findings support the cytological description of both cell lines and general characteristics of more transformed cells having higher sugar and metabolite demands due to higher metabolic rates in the rapid cell cycle and proliferation rates. MCF-7 is compared to MDA-MB-231 less transformed and showed the expected cell

morphology and behavior in culture. Characteristics of slower metabolism of a less transformed cell line could also be shown in the presented results. In contrast MDA-MB-231 exceeds the levels of enrichment of MCF-7 in 14 of 36 individually evaluated metabolites and shows faster enrichment velocities in 9 cases.

The non-targeted search follows the main steps of the targeted approach of data pre-processing (export, overloading correction). As no target library is used peaks need to be detected, aligned and grouped over all sample files in a separate step, using the package *xcms* as described above. After this step, the functions of the package *HiResTEC* are used as described in the *Methods* section to detect peak pairs with a tracer enrichment over time. After peak picking, alignment and testing for peaks having the mass difference of multiples of heavy Carbon atoms (n*1.003355 Da), 7462 peak pairs (mz1/mz2) could be found in the Lymphoma example data set in (Hoffmann et al. 2018), named “preliminary candidate list” (preCL). Of those only 347 candidates, named “evaluated candidate list” (evaCL), passed all filter heuristics provided by the package. The files in supporting information found online for (Hoffmann et al. 2018) contain the peak lists, resulting candidate lists and QC-plots for both, the example data sat and the LC-ESI data set (see below).
Different from other tools, *HiResTEC* does not only rely on the xcms generated peak intensity lists but performs all following calculations and statistical evaluation based on the original raw data. The peak information provided by xcms is used to extract BPCs for the detected m/z-pairs and all corresponding ion traces of the MID from the raw data files. This original spectral and intensity information is the basis of all following calculations of enrichment (changes) and statistical evaluation. Large differences were apparent when using the peak and intensity information from raw data files rather than from the generated xcmsSets. From the initial 7462 peak pairs, 2208 were ranked significantly enriched but only 169 passed the QC filters working on preprocessed data.
Eight such quality filters were implemented and fitted to the specific needs of either GC-APCI or LC-ESI-MS data sets. A specific mass drift filter indicates tracer incorporation, even before tracer enrichment is detectable as an intensity increase of the corresponding isotopologue. Also, it functions to exclude false positives, where the mass drift resulting from tracer accumulation, cannot be observed in the expected fashion. In GC-APCI data this effect was seen in 40 % of the candidates and in 11 % of the candidates in LC-ESI data.

ANOVA models substitute t-tests, which are frequently used in other tools, to evaluate tracer accumulation over time additionally allowing for more factors, like genotypes or treatments, and several experimental labeling time points, thereby increasing flexibility and statistical power. Experiment wide deconvolution extracts mass spectral information for each compound and allows for redundancy removal by detecting peaks and fragments from the same compound in the preCL. Removal of those peaks shortens the list, frees the evaCL from reporting the same compound multiple times, based on a different peak pair and thereby results in less overall computing time. The spectral information is also used for other QC filters that e.g., secure the testing of the base peak of the compound spectrum and general data quality, like existence of the candidate peak pair in raw data or securing that M+0 is being tested.

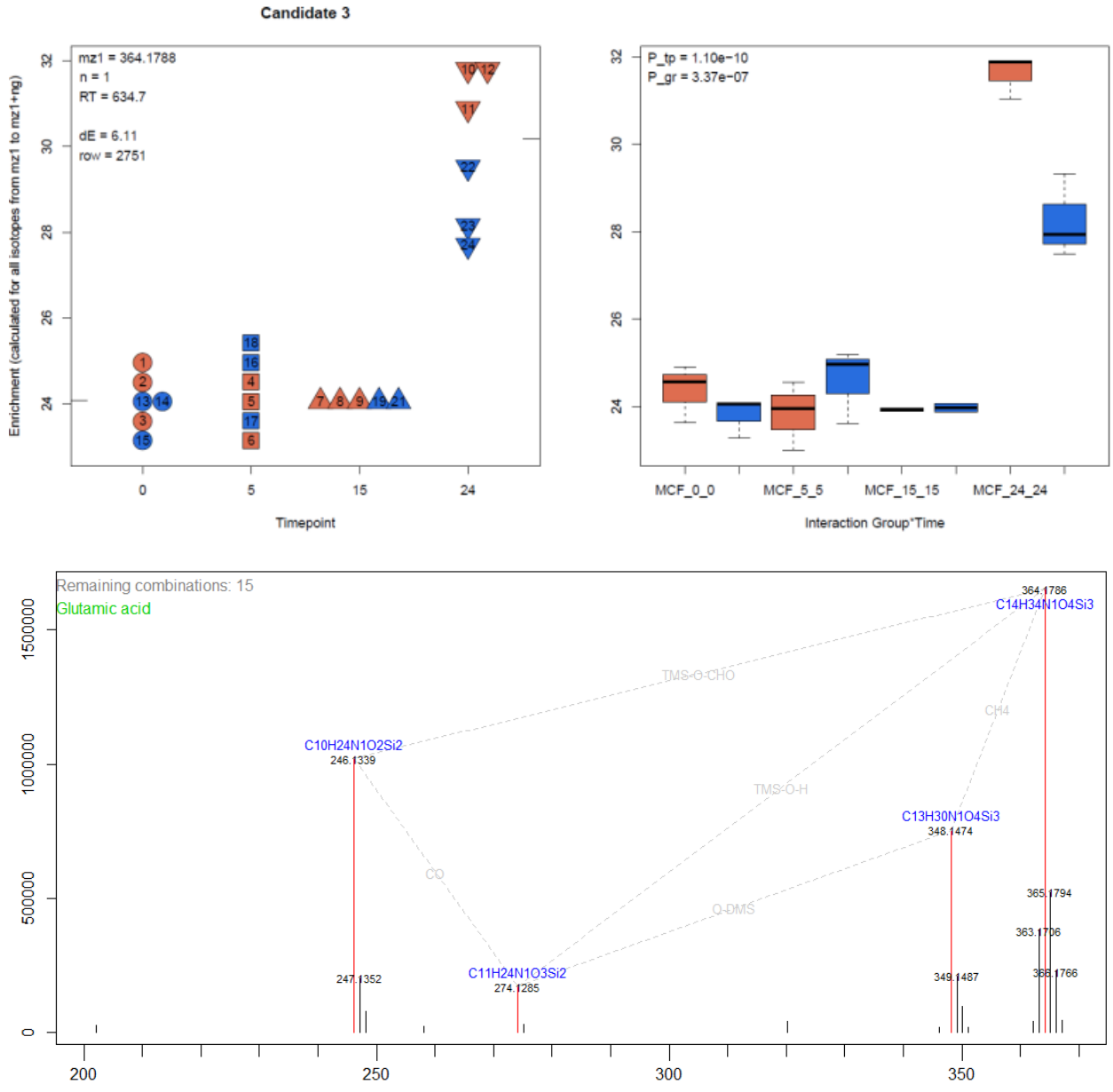
In the breast cancer sample data set 8765 peak pairs (preCL) were found and were subject to the evaluation by the package *HiResTEC*. From those 431 were tested, the remaining pairs were removed from evalCL due to the described spectral overlap/correlation used as redundancy removal. 96 significantly enriched candidates passed all filters and remained in the evalCL; the top 10 lines of the candidate list are shown in Table 3.

*Table 3 Extract of the evaluated candidate table of accepted candidates of the non-targeted search. ID arbitrary number accounting for peak with highest summed intensity and significant enrichment. RT: Retention time of the peak. mz: m/z of peak. ΔE: delta enrichment over time (latest time point). P: time dependent p-value of ANOVA. Name: Name of the compound, assigned after using InterpretMSSpectrum and PubChem search.*

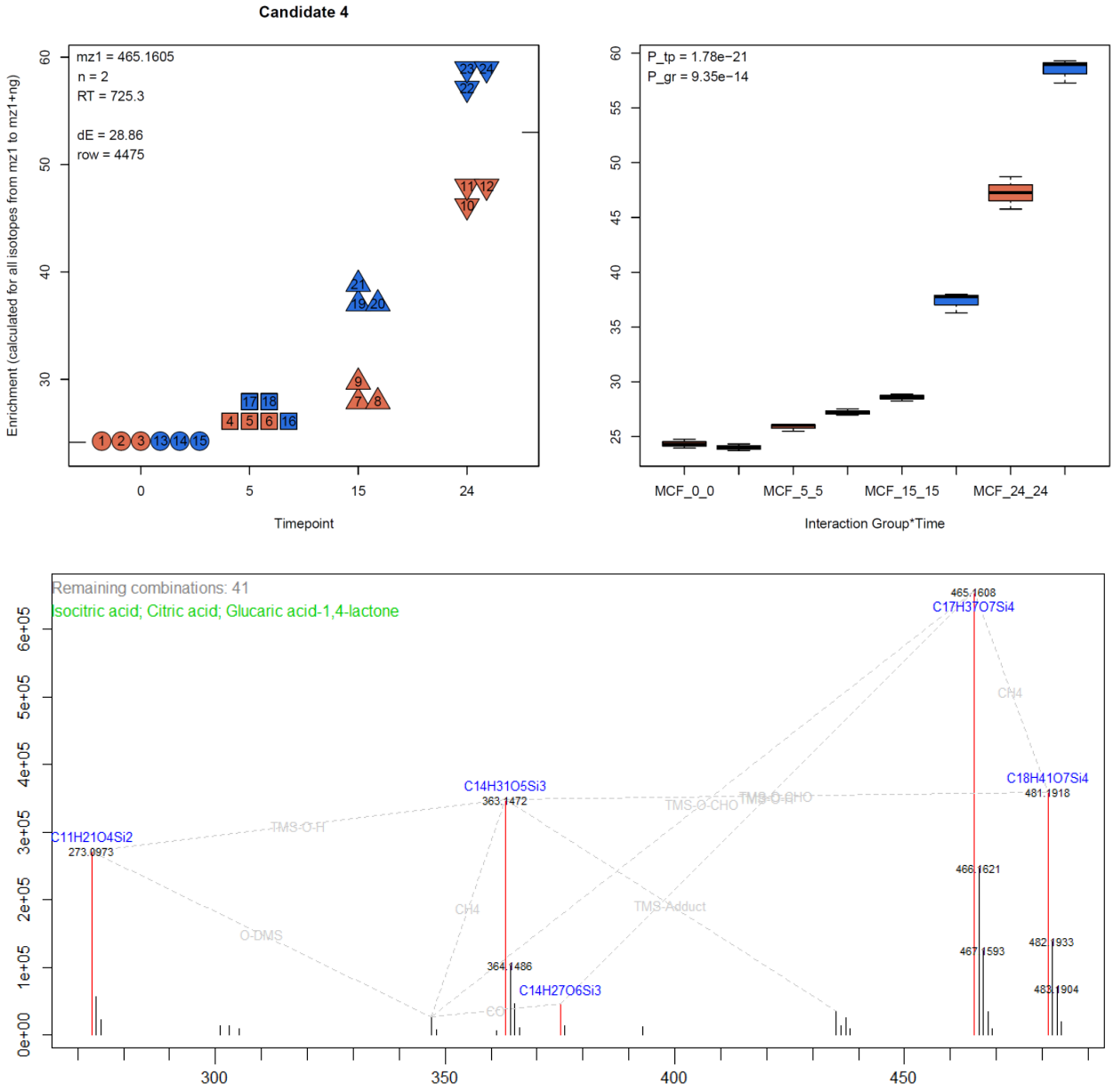
| ID | RT | mz | ΔE | P | Name |
|---|---|---|---|---|---|
| 1 | 762.333 | 319.1574 | 86.79 | 2.41E-13 | C4 labeled Hexose |
| 2 | 770.2585 | 319.1575 | 86.97 | 5.56E-14 | C4 labeled Hexose |
| 3 | 634.6575 | 364.1788 | 6.11 | 1.1E-10 | Glutamic acid |
| 4 | 725.2805 | 465.1605 | 28.86 | 1.78E-21 | Citric acid; Isocitric acid |
| 5 | 352.878 | 234.1341 | 9.03 | 1.79E-16 | Alanine |
| 6 | 325.2505 | 191.0919 | 35.18 | 4.72E-12 | Lactic acid |
| 7 | 588.3935 | 350.1632 | 5.98 | 1.74E-12 | Aspartic acid |
| 8 | 806.113 | 329.2865 | 1.58 | 6.89E-05 | Hexadecanoic acid |
| 9 | 591.2215 | 292.1394 | 5.11 | 1.89E-09 | Serine, O-acetyl- |
| 10 | 590.5405 | 274.1289 | 1.89 | 3.22E-05 | Pyroglutamic acid |

As interpretation of spectra is computation time consuming it is not part of the evaluation procedure but performed after final selection of candidates. This is possible using *InterpretMSSpectrum*, the entries of the GMD and molecular formula search in PubChem for ambiguous sum formula results; the outcome shown here for the first candidates. The identified names of the compounds were added to Table 3. Example plots of the interpreted spectra are shown in Figure 12 A-C. Further plots of the first 10 Candidates are provided in the electronic version.

A



B


Candidate 4
mz1 = 465.1605
n = 2
RT = 725.3
dE = 28.86
row = 4475
Enrichment (calculated for all isotopes from mz1 to mz1+ng)
Timepoint
P_tp = 1.78e−21
P_gr = 9.35e−14
MCF_0_0
MCF_5_5
MCF_15_15
MCF_24_24
Interaction Group*Time
Remaining combinations: 41
Isocitric acid; Citric acid; Glucaric acid-1,4-lactone
C11H21O4Si2
273.0973
C14H31O5Si3
363.1472
C14H27O6Si3
364.1486
C17H37O7Si4
465.1608
C18H41O7Si4
481.1918
466.1621
467.1593
482.1933
483.1904
TMS-O-H
TMS-O-CHO
TMS-O-CHO
CH4
CH4
O-DMS
TMS-Adduct
CO

C

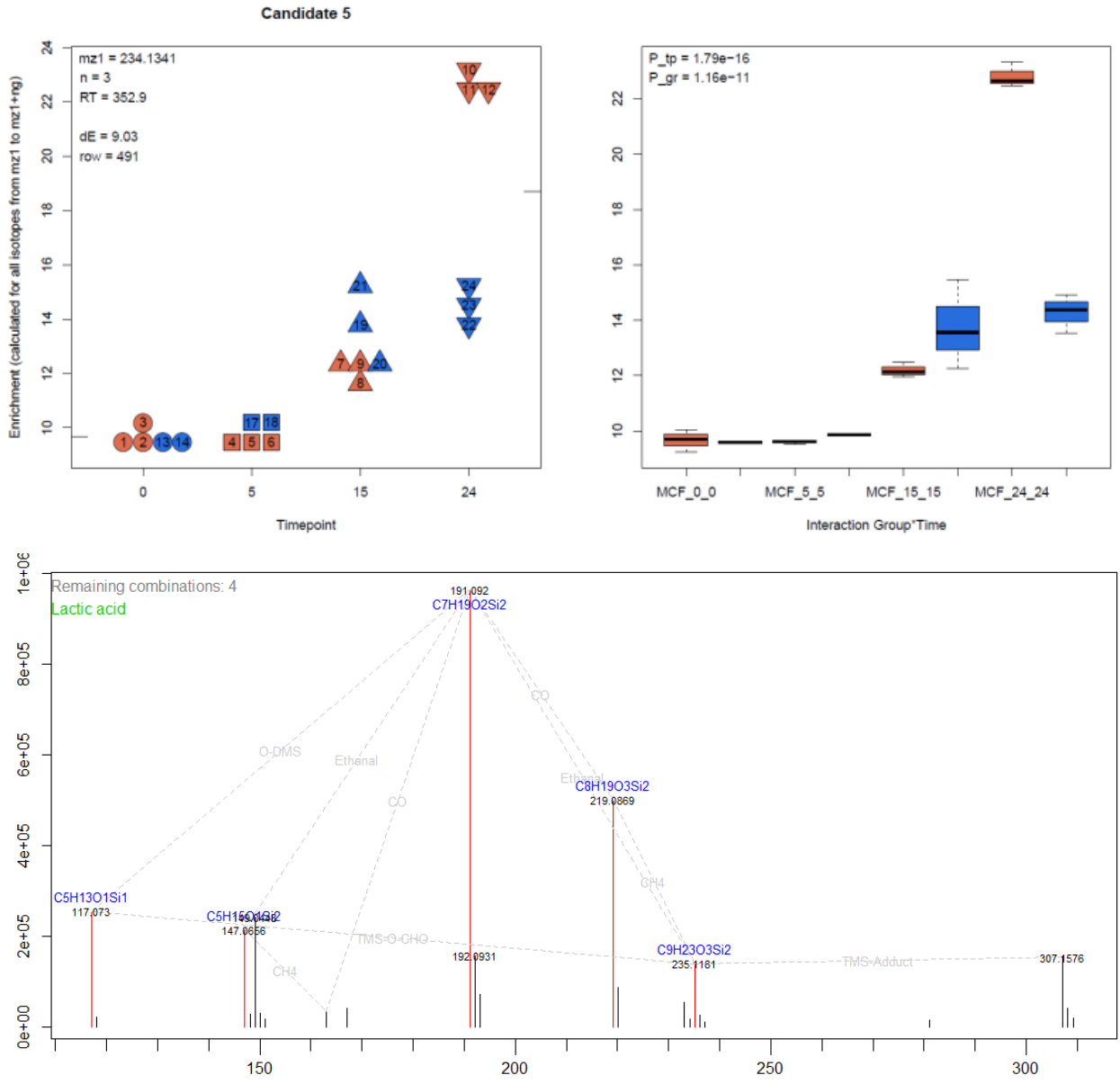



*Figure 12 Selected plots of the non-targeted search, Upper: scatter and box plots with details on the enrichment, statistics and peak under evaluation. Lower: corresponding plots of interpreted spectra. x-axis Intensity; y-axis: m/z. peaks annotated with fragments chemical formula and grey dashed linesgiving information of fragment losses explaining mass differences. A: Candidate 3 (Glutamic acid). B: Candidate 4 (Citric acid). C: Candidate 5 (Lactic acid).*

The evaCL still contains a fraction of false positive hits, which cannot easily be excluded by automatic evaluation. In the Lymphoma test data set the overall FPR was 43 %, and FPR was correlated to peak intensity. In the top 50 candidates no false positive could be found, while in the Top 100 candidates FPR is 11 % and going up for the candidates on the lower end of the list.

Tools designed for LC-MS could not successfully process GC-MS data and resulted in high error rates of diverse origins (Hoffmann et al. 2018). To check the cross-platform functionality of *HiResTEC* a published high-resolution LC-ESI-MS data set was analyzed with two strategies applied. First, the provided raw data files were analyzed with the given data preprocessing parameters, second, the published hits "IncIDs" with their corresponding isotopologues were evaluated with *HiResTEC*. The first approach resulted in a preCL of 20 099 peak pairs to be tested. The reported 271 "IncIDs" resulted in 690 peak pairs to be evaluated. From those, 48 non-redundant candidates could be extracted. The evaluation of the raw data resulted in 53 additional, new candidates, previously not reported by geoRge (Capellades et al. 2016).
The QC plots can be used for fast and easy manual curation and for the tracing of the algorithm's evaluation performance. Thus, false positives rates (FPR) and false negative rates (FNR) were obtained from manual examining QC plots and found FPR=12,5% in 48 mutual candidates and FPR=39,5% in new 53 candidates, FNR was below 3% in both cases for the rejected candidates. This shows, that *HiResTEC* significantly outperforms the previously published fluxomics evaluation tool.
The differences of the two MS-platforms required adapting of the filter heuristics to the individual requirements and features in this data sets. In consequence, different filter heuristics showed to be of different relevance between the data sets. For example, in LC the intensity cut off accounted for 58 % of the rejected candidates, where it has only a very minor relevance in GC-APCI data (0.8 %). The mass drift occurred to a much lesser extent in LC than in GC, 11 % and 40 %, respectively and indicated rather co-eluting compounds than changes due to Si-isotopologues as no derivatization took place. The effect of each QC filter for both data sets and the corresponding cut-offs and assigned error messages are summarized in Table S1 in (Hoffmann et al. 2018).

## Discussion

Metabolites occur in extremely variant concentration ranges in biological samples, which makes it challenging to measure all metabolites in a sample at once. The presented computational solution to correct overloaded signals and estimate their intensity, is a fast and inexpensive option to gain more information in one run.
Signals reaching the detector saturation are flat-topped, i.e. lacking an apex. Those values are detected, removed, and corrected by the algorithm. The validity of the approach could be

shown by artificially cropping peaks and run the algorithms on them. A similar test has been published by (Kalambet et al. 2011), however only for a LC-MS data set and with less relative overloading (up to 5-times above detector saturation, *CorrectOverloadedPeaks*: up to 10-times detector saturation). That study used only one analyte and fewer data points but supports otherwise the main findings and statistical values of *CorrectOverloadedPeaks*.

Besides classical dilution of all samples, or using different GC inlet split ratios, the use of fragments (Wang et al. 2016) or less abundant isotopologues (Trobbiani, Stockham, and Scott 2017) for quantification has been under investigation and provide alternatives to the presented approach. While *CorrectOverloadedPeaks* can be integrated into existing data evaluation pipelines, those methods are more time-consuming and labor intensive, if at all possible, for example, due to limited sample material.

An overloaded signal can occur because of chromatographic overloading, ion suppression in the ion source or due to saturation of the detector. While chromatographic overloading is a rare event in GC-MS measurements; ion suppression is frequent in soft ionization techniques (ESI, APCI), resulting from the limited ionizing capacity of the source. Modeling this process is complex and not subject of the presented R package.

For correction of signals reaching the upper limit of detection due to detector saturation, a computational solution was presented with the stable isotope ratio or Gauss curve approximation in *CorrectOverloadedPeaks*.

Though it is less suited for experiments where the signals have to be precisely quantified, it enables the detection of a larger total number of metabolites per experiment.

Metabolite identification remains a bottleneck in Metabolomics. Especially in non-targeted assays unknown compounds remain largely unidentified. Soft ionization techniques like APCI provided new chances to address this.

Comparison of measured data to reference libraries has been difficult for GC-APCI, as those were only emerging and the coverage was comparatively low (Jaeger et al. 2016).

Spectral matching based annotation often remains inconclusive, even for other MS-based metabolomics set-ups, like LC-ESI. This is often due to technical and biological factors (e.g. instrumentation, biological matrix) causing variation in experimental mass spectra (Jaeger et al. 2017). Thus, manual inspection and interpretation remain a frequent, but non-trivial task and the demand for automation is high.

An open source software package using the information of the preserved molecular ion, typical

neutral losses and a rule-set based on chemical plausibility in metabolites, to assign sum formulas and annotate mass spectra is presented. With a standard mixture of metabolites, it could be shown that the algorithm ranks the correct sum formula on the first three positions in 93 % of the cases. Together with the graphical output this helps the user to evaluate the identity of the compounds much faster.

Exemplary shown on available deconvolution tools (Clasquin, Melamud, and Rabinowitz 2012; Kuhl et al. 2012; Zhang et al. 2014), *InterpretMSSpectrum* proved useful in examining the performance of data pre-processing steps. Differences in peak picking, alignment, and chromatographic deconvolution can lead to bias in large-scale and high-throughput assays. While tools evaluating the peak quality on a single peak basis exist- "zigzag-index"(Zhang and Zhao 2014), "IOP"(Libiseller et al. 2015) *InterpretMSSpectrum* evaluates the spectrum as a whole, and besides annotation can be used also to asses assay quality during analytical method development and optimization.

It is a solution specifically optimized for high-resolution GC-APCI-MS data, taking the conditions of soft ionization and derivatizated analytes into account.

To provide compatibility with LC-ESI methods *InterpretMSSpectrum* was refined and functions were added to determine the precursor in a LC mass spectrum. (Jaeger et al. 2017) This enabled the same functionalities for LC-ESI but taking into account the more complex spectra, different adducts, and neutral losses and the lack of derivatizing agents.

Metabolomics and fluxomics experiments generate significant amounts of data. Exclusive manual examination is thus not feasible, yet the final quality assurance often needs the attention of trained experts in manual labor. *HiResTEC* is a tool that aids the user by reducing the data set to a manageable, assessable size and provides automated QC filters.

To automatize quality checks, first, frequently occurring quality problems and criteria had to be defined. Many of those were found empirically by manual investigation of QC plots and retrospective implementation as a QC filter in the software. Criteria for a good peak are a smooth and Gaussian-like peak shape, a recognizable apex above the baseline, no co-eluting peaks, or if so then good distinguishable (by mass shift, or RT difference), further, for fluxomics, the stable and significant tracer incorporation over time (deltaE and p-values). Peaks were considered less relevant when the tracer enrichment did not surpass 30 % of the median standard deviation of tracer enrichment within replicates or when spectra were sparse and many other peaks from the spectrum were already tested and rejected.

Those criteria were, to a large extend, automatized in *HiResTEC*, thus providing a tool for fast, easy, sensitive, non-redundant, unbiased tracer enrichment calculation in non-targeted fluxomics experiments. The evaluation on the basis of raw data proved to be a very special and valuable tool. As shown above, it leads to the detection of more candidates in total. Also, it uncovers problems with peak picking algorithms (Myers et al. 2017); a higher candidate yield could be expected from peak pairs of the xcmsSet previously rated with a significant tracer enrichment, however, 92% of the candidates were rejected. This high rate could be in part explained by "sensitive peak picking setting, necessary for low abundant isotopologue intensities, [which] can result in peak artifacts in chromatographic noisy regions" (Hoffmann et al. 2018). The top of the list contains peaks with the highest summed intensity secured by the function *RankCandidateList*. Small or spurious peaks, often with lower effects, that are harder to evaluate, are found at the end of the list, explaining the increasing FPR to the lower end. The overall FPR of 43% seems high, however, the algorithm is able to reduce the complexity of the data sets tremendously from a size not manually manageable - several thousands of peaks - to a few hundreds of peaks. In un-targeted fluxomics, the aim is to find possibly all enriched compounds. To demonstrate the performance of the algorithm, cut off values were set rather permissive to report many significantly enriched signals, meeting quality standards, i.e. low false negative rates, on the cost of higher false positive rates. The workload of manual curation on the lower end of the list can be reduced if a fully comprehensive analysis down to the minor signals is not needed, as the most abundant candidates are found on the top of the list with low error rates. While a number of software tools to evaluate $^{13}C$-labeling experiments exist (summarized in Table S3 (Hoffmann et al. 2018)) none was capable to handle comprehensive, non-targeted, high-resolution GC data. Some of the tools were designed with other objectives (iMS2-Flux (Poskar et al. 2012), MetExtract I/II (Bueschl et al. 2012, 2017)), or were only specialized for one technological platform (geoRge (Capellades et al. 2016), x13cms (Huang et al. 2014)) or a lower mass resolution (MIA (Weindl, Wegner, and Hiller 2016) and NTFD (Hiller et al. 2010)). The direct and intensive comparison to one of those tools showed that *HiResTEC* is able to detect more tracer enriched candidates without reporting redundant information like different fragments of the same compound or different combinations of isotopologues of the same MID, and at the same time reporting less false positives due to raw data evaluation.
The development of *HiResTEC* lead to the implementation of a set of functions enabling efficient, sensitive and non-redundant tracer detection and providing possibilities to directly

use the output for flux modeling approaches, by spectral deconvolution, annotation, and MID correction functions. The work of this project provides a coherent pipeline for metabolomic and $^{13}C$-labeled fluxomics data handling, addressing crucial quality control, and (statistical) evaluation steps and methodical bottlenecks of the field.

Streamlined and effective data treatment in metabolomics and fluxomics enhances the (potential) outcome of those experiments, leading to advances in all kinds of biological and medical questions and fields like disease understanding and treatment.

Even though the interest in those fields has increased in the last decades, metabolomic and fluxomic studies remain still heavily underrepresented. Compared to 'genomics' and 'proteomics' Web of Science (Clarivate Analytics 2018) recorded in the year 2017 250-times fewer publications in the field of 'fluxomics' and still two-times fewer publications for 'metabolomics'.

Despite recent advances, fluxomics and metabolomics are still technical, instrumentational and computational demanding experiments, that require expert knowledge and training in diverse disciplines (Rowe, Palsson, and King 2018; Wishart 2016).

The growing community of users and scientists working in the field is more and more equipped with bioinformatic solution and programming skills and provides further solutions for non-expert users and beginners to the field. New tools, like Escher-FBA (Rowe et al. 2018),enable first analysis and even flux modeling steps in easy web applications. Facilitating the step into metabolomics and fluxomics data analysis will lead to more routine use of those technologies and thus adding knowledge to understand complex diseases, like cancer, where just one or two layers of 'omics' are not enough.

Comparability between studies and research in (clinical) metabolomics is still limited. In (His et al. 2019) the common reasons are named: different measurement technologies were applied (e.g. NMR vs. MS, LC, GC, ESI, APCI, EI, TOF, QQQ, …), sample matrix differs, varying materials and experimental procedures have been used, or even in the experiments non-overlapping sets of (target) metabolites are discussed. In that example the researches examine the metabolite levels of blood serum samples to correlate them to breast cancer risk. They discuss openly the limitations of that approach but also point out the future value of such advances.

Measuring metabolites from blood plasma can have a multitude of factors influencing metabolite levels, fasting and non-fasting and circadian rhythms are just the more outstanding variables (Brown 2016). Already at sample retrieval deviation may occur, but also sample

treatment, for example how fast were samples drawn and propagated further to stop enzymatic and chemical reactions to ensure metabolic quenching. Also, metabolite levels alone do not elucidate cellular function. Altered metabolite levels can have various under lying reasons. If it is e.g. higher uptake rate or slower break down rates that leads to elevated levels of a specific metabolite can only be determined by metabolic flux analysis (Martinez-Outschoorn et al. 2017; Sullivan, Gui, and Van Der Heiden 2016).

Despite the fact that the experimental part of this study, has neither the reach nor the coverage of His et al., it nevertheless, shows the potential of metabolomics and fluxomics technologies, already in a small sample set of breast cancer cell lines without any further interventions. The descriptive evaluation of $^{13}C$ labeling experiment of the two breast cancer cell lines could show qualitatively different trends in enrichment and enrichment velocity. It could be observed in accordance with known characteristics of invasive cancer cells, that MDA-MB-231 being a model for such a phenotype, shows a higher enrichment velocity in more metabolites and enriches metabolites higher than the less invasive and transformed cell line MCF-7.

As U-$^{13}C$ -Glucose was used as labeling agent in the experiments, mostly compounds within the sugar metabolizing pathways or in close proximity to these were found enriched in the tracer. An exhaustive analysis of the concerted metabolic interactions cannot be provided with the resources at hand. The untargeted approach is in principle suited to detect and define new target compounds for interventions in cancer therapy. However, a comprehensive testing of these exceeds the scope of this work, which was focused on the proof of concept of the performance of the developed computational solutions and the powerful potential of sensitive metabolic tracer enrichment detection and calculations.

The necessity of such studies in unneglectable as described earlier and in His et al. Careful experimental design and high sample coverage could lead to major steps forward of new discoveries in the field of cancer that can be translated fast to clinical applications.

The diagnostics, prognostics and therapeutics of cancer are in dare need of new pathways. Metabolomics and fluxomics can pave these and help deepen the current understanding of functional interconnections and to tackle the remaining obstacles to cure cancer and other complex diseases.

# References


Aonuma, Masashi, Yoshiyuki Saeki, Toshihiko Akimoto, Yutaka Nakayama, Chiharu Hattori, Yoshino Yoshitake, Katsuzo Nishikawa, Masabumi Shibuya, and Noriko G. Tanaka. 1999. "Vascular Endothelial Growth Factor Overproduced by Tumour Cells Acts Predominantly as a Potent Angiogenic Factor Contributing to Malignant Progression." *International Journal of Experimental Pathology* 80(5):271–81.

Brown, Steven A. 2016. "Circadian Metabolism: From Mechanisms to Metabolomics and Medicine." *Trends in Endocrinology and Metabolism* 27(6):415–26.

Buescher, Joerg M., Maciek R. Antoniewicz, Laszlo G. Boros, Shawn C. Burgess, Henri Brunengraber, Clary B. Clish, Ralph J. DeBerardinis, Olivier Feron, Christian Frezza, Bart Ghesquiere, Eyal Gottlieb, Karsten Hiller, Russell G. Jones, Jurre J. Kamphorst, Richard G. Kibbey, Alec C. Kimmelman, Jason W. Locasale, Sophia Y. Lunt, Oliver Dk Maddocks, Craig Malloy, Christian M. Metallo, Emmanuelle J. Meuillet, Joshua Munger, Katharina Nöh, Joshua D. Rabinowitz, Markus Ralser, Uwe Sauer, Gregory Stephanopoulos, Julie St-Pierre, Daniel a Tennant, Christoph Wittmann, Matthew G. Vander Heiden, Alexei Vazquez, Karen Vousden, Jamey D. Young, Nicola Zamboni, and Sarah-Maria Fendt. 2015. "A Roadmap for Interpreting 13C Metabolite Labeling Patterns from Cells." *Current Opinion in Biotechnology* 34:189–201.

Bueschl, Christoph, Bernhard Kluger, Franz Berthiller, Gerald Lirk, Stephan Winkler, Rudolf Krska, and Rainer Schuhmacher. 2012. "MetExtract: A New Software Tool for the Automated Comprehensive Extraction of Metabolite-Derived LC/MS Signals in Metabolomics Research." *Bioinformatics (Oxford, England)* 28(5):736–38.

Bueschl, Christoph, Bernhard Kluger, Nora K. N. Neumann, Maria Doppler, Valentina Maschietto, Gerhard G. Thallinger, Jacqueline Meng-Reiterer, Rudolf Krska, and Rainer Schuhmacher. 2017. "MetExtract II: A Software Suite for Stable Isotope-Assisted Untargeted Metabolomics." *Analytical Chemistry* 89:9518–26.

Capellades, Jordi, Miriam Navarro, Sara Samino, Marta Garcia-Ramirez, Cristina Hernandez, Rafael Simo, Maria Vinaixa, and Oscar Yanes. 2016. "GeoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics." *Analytical Chemistry* 88(1):621–28.

Carrasco-Pancorbo, Alegría, Ekaterina Nevedomskaya, Thomas Arthen-Engeland, Gabriela Zurek, Carsten Baessmann, André M. Deelder, and Oleg A. Mayboroda. 2009. "Gas Chromatography/ Atmospheric Pressure Chemical Ionization-Time of Flight Mass Spectrometry: Analytical Validation And." *Analytical Chemistry* 81(24):10071–79.

Chavez, Kathryn J., Sireesha V. Garimella, and Stanley Lipkowitz. 2010. "Triple Negative Breast Cancer Cell Lines: One Tool in the Search for Better Treatment of Triple Negative Breast Cancer." *Breast Disease* 32(1–2):35–48.

Clasquin, Michelle F., Eugene Melamud, and Joshua D. Rabinowitz. 2012. "LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine." *Curr Protoc Bioinformatics*.

COMŞA, ŞERBAN, ANCA MARIA CÎMPEAN, and MARIUS RAICA. 2015. "The Story of MCF-7 Breast Cancer Cell Line: 40 Years of Experience in Research." *ANTICANCER RESEARCH* 35(6):3147–54.

DSMZ. n.d. "MCF-7: German Collection of Microorganisms and Cell Cultures GmbH: Details." Retrieved November 27, 2020a (https://www.dsmz.de/collection/catalogue/details/culture/ACC-115).

DSMZ. n.d. "MDA-MB-231: German Collection of Microorganisms and Cell Cultures GmbH: Details." Retrieved November 27, 2020b (https://www.dsmz.de/collection/catalogue/details/culture/ACC-732).

Dunn, Warwick B. 2008. "Current Trends and Future Requirements for the Mass Spectrometric Investigation of Microbial, Mammalian and Plant Metabolomes." *Physical Biology* 5(1):011001.

Dunn, Warwick B., Nigel J. C. Bailey, and Helen E. Johnson. 2005. "Measuring the Metabolome: Current Analytical Technologies." *Analyst* 130(5):606–25.

Dunn, Warwick B., Alexander Erban, Ralf J. M. Weber, Darren J. Creek, Marie Brown, Rainer Breitling,

Thomas Hankemeier, Royston Goodacre, Steffen Neumann, Joachim Kopka, and Mark R. Viant. 2013. “Mass Appeal: Metabolite Identification in Mass Spectrometry-Focused Untargeted Metabolomics.” *Metabolomics* 9(SUPPL.1):44–66.

European Collection of Authenticated Cell Cultures. 2017. *Cell Line Profile MDA-MB-231 (ECACC Catalogue No. 92020424)*.

Fisher, Thomas R., Evelyn B. Haines, and Richard J. Volk. 1979. “A Comment on the Calculation of Atom Percent Enrichment for Stable Isotopes.” *Limnology and Oceanography* 24(3):593–95.

Gest, Caroline, Ulrich Joimel, Limin Huang, Linda Louise Pritchard, Alexandre Petit, Charlène Dulong, Catherine Buquet, Chao Quan Hu, Pezhman Mirshahi, Marc Laurent, Françoise Fauvel-Lafève, Lionel Cazin, Jean Pierre Vannier, He Lu, Jeannette Soria, Hong Li, Rémi Varin, and Claudine Soria. 2013. “Rac3 Induces a Molecular Pathway Triggering Breast Cancer Cell Aggressiveness: Differences in MDA-MB-231 and MCF-7 Breast Cancer Cell Lines.” *BMC Cancer* 13(1):63.

Hanahan, Douglas and Robert A. Weinberg. 2011. “Hallmarks of Cancer: The next Generation.” *Cell* 144(5):646–74.

Hiller, Karsten, Christian M. Metallo, Joanne K. Kelleher, and Gregory Stephanopoulos. 2010. “Nontargeted Elucidation of Metabolic Pathways Using Stable-Isotope Tracers and Mass Spectrometry.” *Analytical Chemistry* 82(15):6621–28.

His, Mathilde, Vivian Viallon, Laure Dossus, Audrey Gicquiau, David Achaintre, Augustin Scalbert, Pietro Ferrari, Isabelle Romieu, N. Charlotte Onland-Moret, Elisabete Weiderpass, Christina C. Dahm, Kim Overvad, Anja Olsen, Anne Tjønneland, Agnès Fournier, Joseph A. Rothwell, Gianluca Severi, Tilman Kühn, Renée T. Fortner, Heiner Boeing, Antonia Trichopoulou, Anna Karakatsani, Georgia Martimianaki, Giovanna Masala, Sabina Sieri, Rosario Tumino, Paolo Vineis, Salvatore Panico, Carla H. Van Gils, Therese H. Nøst, Torkjel M. Sandanger, Guri Skeie, J. Ramón Quirós, Antonio Agudo, Maria Jose Sánchez, Pilar Amiano, José María Huerta, Eva Ardanaz, Julie A. Schmidt, Ruth C. Travis, Elio Riboli, Konstantinos K. Tsilidis, Sofia Christakoudi, Marc J. Gunter, and Sabina Rinaldi. 2019. “Prospective Analysis of Circulating Metabolites and Breast Cancer in EPIC.” *BMC Medicine* 17(1):178.

Hoffmann, Friederike, Carsten Jaeger, Animesh Bhattacharya, Clemens A. Schmitt, and Jan Lisec. 2018. “Nontargeted Identification of Tracer Incorporation in High-Resolution Mass Spectrometry.” *Analytical Chemistry* 90(12):7253–60.

Horning, E. C., M. G. Horning, D. I. Carroll, I. Dzidic, and R. N. Stillwell. 1973. “New Picogram Detection System Based on a Mass Spectrometer with an External Ionization Source at Atmospheric Pressure.” 45(6).

Horning, Evan C., David I. Carroll, I. Dzidic, Klaus D. Haegele, Shen-nan Lin, Christian U. Oertli, Richard N. Stillwell, Ismet Dzidlc, Klaus D. Haegele, Shen-nan Lin, Christian U. Oertli, and Richard N. Stillwell. 1977. “Developmentand Use of Analytical Systems Based on Mass Spectrometry.” *Clinical Chemistry* 23(1):13–21.

Huang, Xiaojing, Ying-Jr Chen, Kevin Cho, Igor Nikolskiy, Peter A. Crawford, and Gary J. Patti. 2014. “X13CMS: Global Tracking of Isotopic Labels in Untargeted Metabolomics.” *Analytical Chemistry* 86:1632−1639.

Jaeger, Carsten, Friederike Hoffmann, Clemens A. Schmitt, and Jan Lisec. 2016. “Automated Annotation and Evaluation of In-Source Mass Spectra in GC/Atmospheric Pressure Chemical Ionization-MS-Based Metabolomics.” *Analytical Chemistry* 88(19):9386–90.

Jaeger, Carsten, Michaël Méret, Clemens A. Schmitt, and Jan Lisec. 2017. “Compound Annotation in Liquid Chromatography/High-Resolution Mass Spectrometry Based Metabolomics: Robust Adduct Ion Determination as a Prerequisite to Structure Prediction in Electrospray Ionization Mass Spectra.” *Rapid Communications in Mass Spectrometry* 31(15):1261–66.

Kalambet, Yuri, Yuri Kozmin, Ksenia Mikhailova, Igor Nagaev, and Pavel Tikhonov. 2011. “Reconstruction of Chromatographic Peaks Using the Exponentially Modified Gaussian Function.” *Journal of Chemometrics* 25(7):352–56.

Kopka, J., N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser. 2005. “GMD@CSB.DB: The Golm

Metabolome Database." *Bioinformatics* 21(8):1635–38.
Kuhl, Carsten, Ralf Tautenhahn, Christoph B?ttcher, Tony R. Larson, Steffen Neumann, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, and Christoph Magnes. 2012. "CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets." *Analytical Chemistry* 84(1):283–89.
Libiseller, Gunnar, Michaela Dvorzak, Ulrike Kleb, Edgar Gander, Tobias Eisenberg, Frank Madeo, Steffen Neumann, Gert Trausinger, Frank Sinner, Thomas Pieber, and Christoph Magnes. 2015. "IPO: A Tool for Automated Optimization of XCMS Parameters." *BMC Bioinformatics* 16(1):118.
Lisec, Jan, Friederike Hoffmann, Clemens Schmitt, and Carsten Jaeger. 2016. "Extending the Dynamic Range in Metabolomics Experiments by Automatic Correction of Peaks Exceeding the Detection Limit." *Analytical Chemistry* 88(15):7487–92.
Martinez-Outschoorn, Ubaldo E., Maria Peiris-Pagés, Richard G. Pestell, Federica Sotgia, and Michael P. Lisanti. 2017. "Cancer Metabolism: A Therapeutic Perspective." *Nature Reviews Clinical Oncology* 14(1):11–31.
Moreno-Sánchez, Rafael, Emma Saavedra, Juan Carlos Gallardo-Pérez, Franklin D. Rumjanek, and Sara Rodríguez-Enríquez. 2016. "Understanding the Cancer Cell Phenotype beyond the Limitations of Current Omics Analyses." *FEBS Journal* 283(1):54–73.
Myers, Owen D., Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du. 2017. "Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data." *Analytical Chemistry* 89:8689−8695.
Nowicki, Stefan and Eyal Gottlieb. 2015. "Oncometabolites: Tailoring Our Genes." *FEBS Journal* 282(15):2796–2805.
Otto Warburg, Karl Posener, and Erwin Negelein. 1924. "Über Den Stoffwechsel Der Carcinomzelle." *Z Biochemie* 152:309–44.
Poskar, C. Hart, Jan Huege, Christian Krach, Mathias Franke, Yair Shachar-Hill, and Björn H. Junker. 2012. "IMS2Flux--a High-Throughput Processing Tool for Stable Isotope Labeled Mass Spectrometric Data Used for Metabolic Flux Analysis." *BMC Bioinformatics* 13(1):295.
R Core Team. 2017. "R: A Language and Environment for Statistical Computing."
Rowe, Elliot, Bernhard O. Palsson, and Zachary A. King. 2018. "Escher-FBA: A Web Application for Interactive Flux Balance Analysis." *BMC Systems Biology* 12(1):84.
Sauer, Uwe. 2006. "Metabolic Networks in Motion: 13 C-Based Flux Analysis." *Molecular Systems Biology* 2(1).
Shirazi, Farshad H., Afshin Zarghi, Farzad Kobarfard, Rezvan Zendehdel, Maryam Nakhjavani, Sara Arfaiee, Tannaz Zebardast, Shohreh Mohebi, Nassim Anjidani, Azadeh Ashtarinezhad, and Shahram Shoeibi. 2011. "Remarks in Successful Cellular Investigations for Fighting Breast Cancer Using Novel Synthetic Compounds." Pp. 85–102 in *Breast Cancer - Focusing Tumor Microenvironment, Stem cells and Metastasis*. InTech.
Smith, C. A., E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak. 2006. "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification." *Anal Chem* 78:779–87.
Strehmel, Nadine, Joachim Kopka, Dierk Scheel, and Christoph Böttcher. 2014. "Annotating Unknown Components from GC/EI-MS-Based Metabolite Profiling Experiments Using GC/APCI(+)-QTOFMS." *Metabolomics* 10(2):324–36.
Sullivan, Lucas B., Dan Y. Gui, and Matthew G. Van Der Heiden. 2016. "Altered Metabolite Levels in Cancer: Implications for Tumour Biology and Cancer Therapy." *Nature Reviews Cancer* 16(11):680–93.
Trobbiani, Stephen, Peter Stockham, and Timothy Scott. 2017. "Increasing the Linear Dynamic Range in LC-MS: Is It Valid to Use a Less Abundant Isotopologue?" *Drug Testing and Analysis* 9(10):1630–36.
Tsugawa, Hiroshi, Takeshi Bamba, Masakazu Shinohara, Shin Nishiumi, Masaru Yoshida, and Eiichiro Fukusaki. 2011. "Practical Non-Targeted Gas Chromatography/Mass Spectrometry-Based Metabolomics Platform for Metabolic Phenotype Analysis." *Journal of Bioscience and*

*Bioengineering* 112(3):292–98.
Wachsmuth, Christian J., Martin F. Almstetter, Magdalena C. Waldhier, Michael A. Gruber, N. Nadine, Peter J. Oefner, and Katja Dettmer. 2011. "Performance Evaluation of Gas Chromatography-Atmospheric Pressure Chemical Ionization-Time-of-Flight Mass Spectrometry for Metabolic Fingerprinting and Profiling." *Analytical Chemistry* 7514–22.
Wachsmuth, Christian J., Thomas A. Hahn, Peter J. Oefner, and Katja Dettmer. 2015. "Enhanced Metabolite Profiling Using a Redesigned Atmospheric Pressure Chemical Ionization Source for Gas Chromatography Coupled to High-Resolution Time-of-Flight Mass Spectrometry." *Analytical and Bioanalytical Chemistry* 407(22):6669–80.
Wang, Yi, Haiyan Hu, Yue Su, Fang Zhang, and Yinlong Guo. 2016. "Potential of Monitoring Isotopologues by Quantitative Gas Chromatography with Time-of-Flight Mass Spectrometry for Metabolomic Assay." *Journal of Separation Science*.
Weindl, Daniel, Thekla Cordes, Nadia Battello, Sean C. Sapcariu, Xiangyi Dong, Andre Wegner, and Karsten Hiller. 2016. "Bridging the Gap between Non-Targeted Stable Isotope Labeling and Metabolic Flux Analysis." *Cancer & Metabolism* 4(1):10.
Weindl, Daniel, Andre Wegner, and Karsten Hiller. 2016. "MIA: Non-Targeted Mass Isotopolome Analysis." *Bioinformatics (Oxford, England)* btw317.
Welsh, Jo Ellen. 2013. "Animal Models for Studying Prevention and Treatment of Breast Cancer." Pp. 997–1018 in *Animal Models for the Study of Human Disease*. Elsevier Inc.
Wiechert, W. 2001. "13C Metabolic Flux Analysis." *Metabolic Engineering* 3(3):195–206.
Wiechert, Wolfgang, Michael Mo, So Petersen, and Albert a De Graaf. 2001. "A Universal Framework for 13C Metabolic Flux Analysis." *Metabolic Engineering* 3(3):265–83.
Winter, Gal and Jens O. Krömer. 2013. "Fluxomics - Connecting 'omics Analysis and Phenotypes." *Environmental Microbiology* 15(7):1901–16.
Wishart, David S. 2016. "Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine." *Nature Reviews Drug Discovery* 15(7):473–84.
Wishart, David S., Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. 2013. "HMDB 3.0—The Human Metabolome Database in 2013." *Nucleic Acids Research* 41(D1):D801–7.
Zamboni, Nicola. 2011. "13C Metabolic Flux Analysis in Complex Systems." *Current Opinion in Biotechnology* 22(1):103–8.
Zamboni, Nicola, Sarah-Maria Fendt, Martin Rühl, and Uwe Sauer. 2009. "13C-Based Metabolic Flux Analysis." *Nature Protocols* 4(6):878–92.
Zhang, Wenchao, Junil Chang, Zhentian Lei, David Huhman, Lloyd W. Sumner, and Patrick Xuechun Zhao. 2014. "MET-COFEA: A Novel Liquid Chromatography-Mass Spectrometry Data Processing Platform for Metabolite Compound Feature Extraction and Annotation." *Analytical Chemistry*.
Zhang, Wenchao and Patrick X. Zhao. 2014. "Quality Evaluation of Extracted Ion Chromatograms and Chromatographic Peaks in Liquid Chromatography/Mass Spectrometry-Based Metabolomics Data." *BMC Bioinformatics* 15(Suppl 11):S5.

# Abkürzungsverzeichnis

| | |
|---|---|
| ANOVA | Analysis of variance |
| APCI | Atmospheric pressure chemical ionization |
| BPC | Base Peak Chromatogram |
| CRAN | The Comprehensive R Archive Network |
| DMEM | Dubleco´s Modified Essential Medium |
| EI | Electron impact (ionisation) |
| ESI | Electron spray ionisation |
| evaCL | Evaluated Candidate List |
| EWD | Experiment Wide Deconvolution |
| FBS | Fetal Bovine Serum |
| FNR | False Negative Rate |
| FPR | False Positive Rates |
| G | Gauss |
| GC | Gas Chromatography |
| GMD | Golm Metabolomic Database |
| HiResTec | High-resolution Tracer Enrichment Calculation |
| HMDB | Human Metabolome Database |
| IR | Isotopic ratio |
| LC | Liquid Chromatography |
| LR | Linear range |
| M+0 | Molecular peak (mass +0 refers to no heavy isotopes are incorporated) |
| MDV | Mass distribution vector |
| MID | Mass Isotopomer Distribution |
| MS | Mass spectrometry |
| MSTFA | N-Methyl-N-(trimethylsilyl) trifluoroacetamide |
| O-DMS | O-Dimethylsulfide |
| preCL | Preliminary Candidate List |
| QC | Quality Control |
| R | Open source statistics scripting language |
| RT | Retention Time |
| TMS-OH | Trimethylsilyl-OH |
| TOF | Time of light |

# Anteilserklärung

**Eidesstattliche Versicherung**

„Ich, Friederike Hoffmann, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema:
„Analysis of metabolic fluxes in cancer cell lines with novel computational tools to enable the non-targeted tracer incorporation detection from high-resolution mass spectrometry data“
(deutsch: „Analyse von Stoffwechselflüssen in Krebszelllinien mit neuartigen computergestützten Werkzeugen, um die nicht-zielgerichtete Tracer-Inkorporation aus hochauflösenden Massenspektrometriedaten zu ermöglichen“) selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Erstbetreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

_____

Datum Unterschrift

**Anteilserklärung an den erfolgten Publikationen**

Friederike Hoffmann hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1:

Lisec, Jan, Friederike Hoffmann, Clemens Schmitt, and Carsten Jaeger,
*'Extending the Dynamic Range in Metabolomics Experiments by Automatic Correction of Peaks Exceeding the Detection Limit'*,
Analytical Chemistry, (2016)

Beitrag im Einzelnen:

An der Erstellung des Manuskripts habe ich unterstützend mitgewirkt und meine kritische Begutachtung beigetragen. Die Literaturrecherche wurde von mir durchgeführt und vervollständigt, sowie die korrekten Zitiermethode und Verweise eingefügt.
Die Datenauswertung und abschließende Test habe ich in Teilen durchgeführt. Dazu war die funktionelle Evaluation des Algorithmus und das Testen an verschiedene vorliegenden Datensätzen notwendig, sowie Installation auf verschieden Geräten und Plattformen (Windows und Linux).
Im Verlauf habe ich dazu nach der Analyse und der Revision der Funktionen des Software-Pakets die Abbildung Figure 1 vollständig selbständig erstellt und mich bei Figure 2 zur Ausgestaltung der zugrundeliegenden Daten in graphischer Form eingesetzt und dem Layout mitgewirkt. Die übrigen Abbildungen im Manuscript sind aus vorangegangen Datenanalysen durch Dr. Jan Lisec und Dr. Carsten Jaeger entwickelt worden, beziehungsweise Figure 3 aus den durch den Algorithmus automatisch generieten Plots.
Im Revisionsprozess des Manuskripts war ich vollständig eingebunden und habe einige Gutachterfragen adressiert und bei der Einarbeitung der Anmerkungen mitgewirkt.

Publikation 2:

Jaeger, Carsten, Friederike Hoffmann, Clemens A. Schmitt, and Jan Lisec,
*'Automated Annotation and Evaluation of In-Source Mass Spectra in GC/Atmospheric Pressure Chemical Ionization-MS-Based Metabolomics'*,
Analytical Chemistry, 88 (2016)

Beitrag im Einzelnen:

An der Erstellung des Manuskripts habe ich unterstützend mitgewirkt und meine kritische Begutachtung beigetragen. Die Literaturrecherche wurde von mir durchgeführt und vervollständigt, die korrekten Zitiermethode und Verweise eingefügt.
Die im Manuskript beschriebenen Methoden wurden von mir etabliert. Namentlich, für die Massenspektrometrie geeignete Probenvorbereitung und die verschiedenen Extraktionsmethoden aus biologischer Matrix auf Perfomanzunterschiede hin untersucht und vollständig durchgeführt, und auf die erwähnten zellbiologischen Proben angewandt.
Bei der programmatischen Erstellung der Scripte des Algorithmus habe ich in Teilen mitgewirkt. Die anschließende Datenauswertung konnte ich in Teilen mit durchführen und abschließende Test durchführen.
Dazu war die funktionelle Evaluation des Algorithmus und das Testen an verschiedene vorliegenden Datensätzen notwendig, sowie Installation auf verschieden Geräten und Plattformen (Windows und Linux). Nach der Analyse und der Revision der Funktionen des Software- Pakets Figure S-1 wurde von mir vollständig, selbstständig erstellt.

Publikation 3:

Hoffmann, Friederike, Carsten Jaeger, Animesh Bhattacharya, Clemens A. Schmitt, and Jan Lisec, 'Nontargeted Identification of Tracer Incorporation in High-Resolution Mass *Spectrometry'*, Analytical Chemistry, 90 (2018)

Beitrag im Einzelnen:

Die funktionelle Entwicklung der Konzeption erfolgte vollständig und selbständig von mir und in Rücksprache mit Dr. Jan Lisec. Die Literaturrecherche und -auswertung wurde vollständig von mir durchgeführt. Die Methodenentwicklung gliederte sich in verschiedene Bereiche, ebenso das Versuchsdesign, die zum weitaus überwiegenden Anteil von mir selbst angefertigt wurden. Die praktischen Arbeiten und Planung für alle im biologischen Labor anfallenden Arbeiten habe ich selbst und vollständig übernommen. Namentlich, Ablaufpläne, Materialbestellung, und -verwaltungen, Anzucht von Krebszellkulturen, Etablierung und Durchführung von 13C-Tracer-Labeling Experimenten, Zellaufschluss und Extraktion der Metabolite. Die Planung und Optimierung der analytischen Mess-Methoden wurden unterstützt durch Dr. Jan Lisec und Dr. Carsten Jaeger. Die Vorbereitung der Messungen am GC-MS (Derivatisierung) und die Messung wurden vollständig von mir selbst durchgeführt. Die Programmierung der Scripte des Algorithmus wurde von mir in Teilen selbst durchgeführt und zu Teilen durch Dr. Jan Lisec deutlich unterstützt. Da einige der Scripte erheblich in die Daten-Vorprozessierung eingreifen und ausgiebiges massenspektrometrisches Expertenwissen vorrausetzen wurden ich auch bei der Roh-Daten-Handhabung zum Teil von Dr. Jan Lisec unterstützt. Die Evaluierung und funktionelle Analyse aller Scripte, der kritische Vergleich und die Gegenüberstellung wurde vollständig von mit selbst durchgeführt.
Die Datenauswertung der biologischen Daten mit den programmatisch entwickelten Scripten wurde vollständig von mir selbst durchgeführt.
Nach Abschluss der Entwicklung aller Funktionen der Softwarepakets ist die Abbildung Figure 1 selbstständig und vollständig von mir erstellt worden.

Des Weiteren, sind im Detail die Abbildung Figure 3 nach kritischer Auseinandersetzung mit existierender Software und teils manueller Revision und Vergleich der Leistungsunterschiede der Algorithmen sowie die Tabelle Table S3, die die Datenhandhabung und Auswertungsfunktionalität von publizierten Software-Tools umfassend darstellt, vollständig und selbständig von mir erstellt worden.
Die Abbildung Figure 2 ist Produkt der automatisierten Auswertungs-Scripte und wurde mit hilfreichen Anmerkungen versehen. Die Tabelle Table S1 entstand aus zum Teil von mir selbst nach statistischer Analyse gewonnen Daten und ist von Dr. Jan Lisec mit weiteren Details und Daten einer Vergleichsanalyse erweitert worden.
Das Erstellen des Textkörpers des Manuskripts, sowie die kritische Auseinandersetzung mit den gewonnenen Daten zur Leistungsfähigkeit der Scripte und Anwendbarkeit und alle dazugehörigen Schritte im Publikationsprozess wurden von mir selbst durchgeführt.
Während des Revisionsprozess des Manuskripts wurde ich von Dr. Jan Lisec und Dr. Carsten Jaeger hilfreich unterstützt, und habe alle wesentlichen Schritte und Leistungen, wie Adressierung von Gutachterfragen und -anmerkungen, Punkt-zu-Punkt-Beantwortung, Überarbeitung, Verfassen von Anschreiben usw. selbst erbracht.

_____

Unterschrift, Datum und Stempel des betreuenden Hochschullehrers

_____

Unterschrift der Doktorandin

Publikationen

# Extending the Dynamic Range in Metabolomics Experiments by Automatic Correction of Peaks Exceeding the Detection Limit



analytical chemistry

Technical Note

pubs.acs.org/ac

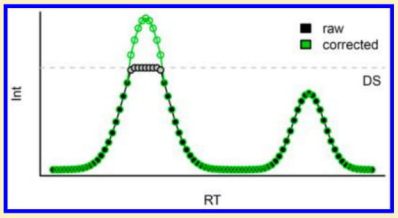
## Extending the Dynamic Range in Metabolomics Experiments by Automatic Correction of Peaks Exceeding the Detection Limit

Jan Lisec,[*,†,‡] Friederike Hoffmann,[†] Clemens Schmitt,[†,§,∥] and Carsten Jaeger[†,∥]

[†]Charité-Universitätsmedizin Berlin, Molekulares Krebsforschungszentrum (MKFZ), Augustenburger Platz 1, 13353 Berlin, Germany
[‡]German Cancer Consortium, Deutsches Krebsforschungzentrum (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[§]Max-Delbrück-Center for Molecular Medicine (MDC), Robert-Rössle-Straße 10, 13125 Berlin, Germany
[∥]Berlin Institute of Health (BIH), Kapelle-Ufer 2, 10117 Berlin, Germany

Supporting Information

**ABSTRACT:** Metabolomics, the analysis of potentially all small molecules within a biological system, has become a valuable tool for biomarker identification and the elucidation of biological processes. While metabolites are often present in complex mixtures at extremely different concentrations, the dynamic range of available analytical methods to capture this variance is generally limited. Here, we show that gas chromatography coupled to atmospheric pressure chemical ionization mass spectrometry (GC-APCI-MS), a state of the art analytical technology applied in metabolomics analyses, shows an average linear range (LR) of 2.39 orders of magnitude for a set of 62 metabolites from a representative compound mixture. We further developed a computational tool to extend this dynamic range on average by more than 1 order of magnitude, demonstrated with a dilution series of the compound mixture, using robust and automatic reconstruction of intensity values exceeding the detection limit. The tool is freely available as an R package (CorrectOverloadedPeaks) from CRAN (https://cran.r-project.org/) and can be incorporated in a metabolomics data processing pipeline facilitating large screening assays.






Metabolomics aims to detect, identify, and quantify all small molecules present in a biological sample to ultimately gain insight into the functionality of biological processes or identify biomarkers, i.e., metabolites whose levels are indicative of disease or some other status of the analyzed biological system.[1−3] Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are the two technologies most widely used to detect and differentiate between metabolites, where the latter is often coupled to a preceding separation step, mostly gas chromatography (GC), liquid chromatography (LC), or capillary electrophoreses (CE), to facilitate detection.[4]

While current analytical systems are capable of distinguishing several thousand signals within a sample, quantifying these metabolites is challenging due to their chemical complexity and large dynamic concentration range. According to the Human Metabolome Database,[5] metabolite concentrations in human plasma and urine vary from the pico- up to the millimolar range (Figure S1A). Within experiments, individual metabolites are mostly reported to vary less than 2-fold (Figure S1B), but experiments where metabolite levels vary more than 3 orders of magnitude are reported as well. Large concentration ranges and chemical complexity may be accounted for by parallel analysis of samples in dilution series and on different analytical platforms (GC/LC−MS, NMR). In practice, however, material, originating from biopsies or primary cell culture assays, is often limited and analysis time is costly.

Regarding mass spectrometry, detection limits of modern time-of-flight (TOF) or Orbitrap instruments used in metabolomics screening assays are specified at ∼100 pM at signal-to-noise ratios (S/N) of 100:1 at least for specific test substances such as reserpine. Increasing sensitivity allows to detect even minor compounds as potential biomarkers in small sample amounts and to scale down processing volumes and, therefore, decrease cost of chemicals which is of relevance in large scale experiments.

As a drawback, decreasing the lower limit of detection in sensitive assays usually leads to highly abundant signals in complex mixtures exceeding the detection limit of the analytical system. Often these compounds are simply removed from downstream analysis, evaluated separately by measuring a diluted sample series or approximated based on isotope/fragment information if available. Alternatively, attempts have been made to computationally reconstruct the signal.[6]

In GC-APCI-MS data intensity values exceeding digitizer saturation (IVEDS) are common in analyses of complex biological samples. Often these peaks are still symmetric. We here implemented two algorithms, based on isotopic ratio and Gaussian peak shape, to achieve a robust reconstruction of

**Received:** July 1, 2016
**Accepted:** July 5, 2016
**Published:** July 5, 2016





DOI: 10.1021/acs.analchem.6b02515
*Anal. Chem.* 2016, 88, 7487−7492

IVEDS. We show in dilution series of complex but defined chemical mixtures that it is possible to correct peaks up to 10-fold higher in concentration than the upper limit of detection of the measurement system within the error of the linear range, thereby increasing the dynamic range by 1 order of magnitude or 50%, respectively, in an automatic fashion. We describe the algorithm, which is provided as an R package working on mzXML data files and hence can be incorporated into any metabolomics processing pipeline as a preprocessing step. We further discuss limitations of this approach using quality control plots generated by the software.

## EXPERIMENTAL SECTION

**Sample Preparation, Derivatization, and GC-APCI-MS Analysis.** A total of 62 metabolites were purchased as chemical standards from Sigma-Aldrich (Germany) in reagent-grade quality (see Table S1), combined in a master mix at 1.25 mM, serially diluted in 24 steps to 12.8 pM and subjected to derivatization and GC-APCI-MS measurement as described in more detail in Methods S1.

**Peak Correction Function "CorrectOverloadedPeaks".** An algorithm for automated reconstruction of peaks exceeding the detection limit in APCI-MS was implemented as an R (https://www.r-project.org/) function and can be installed as a package (CorrectOverloadedPeaks) from CRAN (https://cran.r-project.org/). For convenience, the function accepts data stored as an xcmsRaw object, which can be generated from various file formats using the freely available xcms-package.[7] As a result, peaks within this xcmsRaw object exceeding a certain limit will be extrapolated using either of two methods (isotopic ratio or Gaussian approach, see further below) and stored back in the xcmsRaw object for further processing. As an alternative, mzXML files can be processed directly.

The intensity limit $i$ separating values to be corrected from values used for this correction is calculated by $i = n \times DS$, where $n$ is a weighting factor and DS is digitizer saturation. DS is a discrete value which is dependent on the number of TOF events per second (which itself is dependent on the mass range covered, as larger masses require more time to flight and hence decrease the number of TOF events) and the digitizer capacity (here 10 bit or 1024 units). In our setup (10 Hz, mass range 50−1000) we can record 949 TOF events per scan which translates to $DS = 949 \times 1024 - 1 = 971\,775$ counts, i.e., 971 775 ions of a certain mass which can be counted at the max within a single scan. Measured intensity values equal to DS indicate that we reached or exceeded the upper detection limit of our analytical system and all raw data values equivalent to DS should hence be corrected. However, also values approaching DS should potentially be corrected. We set $n = 0.95$ for this purpose but would not recommend to specify values <0.85 because this will limit the amount of data points used to reconstruct (as values equal to or larger $i$ are considered wrong and will be substituted based on values of smaller $i$).

The algorithm (Figure 1) incorporates the following processing steps. (1) If not specified explicitly by the user as a parameter, DS is determined from the supplied data by a heuristic analysis of the base peak chromatogram. (2) All chromatographic areas containing IVEDS (peaks higher than $i = n \times DS$) are detected. We specified $n = 0.95$. (3) For all ion traces with values larger than $i$ base peak chromatograms (BPCs) in narrow mass windows are extracted around the respective region. (4) From these BPCs all scan values larger than $i$ will be modified using either (i) isotopic ratio (IR) information or (ii) using an assumption of Gaussian peak shape or exponentially modified Gaussian peak shape and modeling IVEDS based on front and tail values of the peak. In both cases, quality control plots will be generated for every modified ion trace on demand. (5) The modified data are returned as xcmsRaw or mzXML, depending on the input.

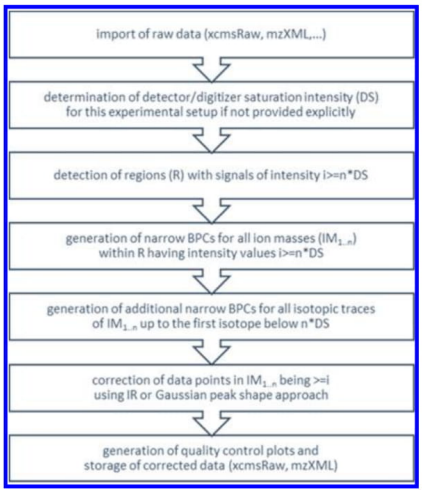



**Figure 1.** Individual processing steps of the CorrectOverloadedPeaks function.

The IR approach, more specifically, works by finding the first isotopic peak not higher than $i$ itself, extracting its BPC, testing for the stable ratio between the peak to correct and the isotopic peak based on nonoverloaded values in the peak front/tail, and modeling IVEDS by scaling with the observed isotopic ratio. In example, if for an ion of $m/z = 100$ five scans with intensity = $i$ are detected at a retention time 300 s, than three narrow BPCs are extracted for M ($m/z = 100$), M+1 ($m/z = 101.003$), and M +2 ($m/z = 102.006$), all at 300 ± 2 s. Values in similar scans below $i$ are used to estimate the stable ratios $r_1$ (M+1/M) and $r_2$ (M+2/M), respectively. If all values of M+2 are below $i$, we can transform all values of M larger than $i$ using $M' = M+2/r_2$.

Fronting and/or tailing of a peak is determined empirically by calculating the ratio between baseline levels before and after the peak. If this ratio is found to be outside the range of 0.2−5 than either fronting or tailing are assumed and error terms during gauss correction are neglected for the distorted side of the peak.

**Determination of Linear Ranges.** Calculation of limits of detection (LOD), quantification (LOQ), and linearity (LOL) were adapted from Konieczka and Namiesnik.[8] Initial values for LOD and LOQ ($LOD_{init}$, $LOQ_{init}$) were estimated from six blank measurements (LOD = 3 × median peak intensity, LOQ = 6 × median peak intensity). Next, more precise estimates of LOD and LOQ were obtained by fitting a linear model through the peak intensities of low-concentration measurements (intensity < $LOQ_{init}$) using equations $LOD = 3.3s/m$ and $LOQ = 10s/m$ where $s$ is the standard deviation of the residuals and $m$ the slope of the linear fit. For LOL determination, all points above LOQ were subjected to piecewise linear


DOI: 10.1021/acs.analchem.6b02515
Anal. Chem. 2016, 88, 7487−7492

regression analysis as implemented by R package SiZer,[9] using the statistical model:

$$Y = \beta_0 + \beta_1(X) + \beta_2(X - C)^+ + \epsilon \quad (1)$$

where $Y$ is the peak intensity, $X$ is the concentration, $C$ is the changepoint (i.e., the LOL), and $\beta_0$, $\beta_1$, and $\beta_2$ the intercept, change in slope prior to the changepoint, and change in slope after the changepoint to be estimated, respectively. Fitting this model to the calibration curve resulted in identification of a first and (possibly) second linear portion of the curve. The first linear portion corresponded to the linear calibration range, with its end, the changepoint, marking the LOL. The linear range (LR) was finally calculated as $\log_{10}$(LOL/LOQ).

## RESULTS AND DISCUSSION

Extreme dynamic concentration ranges of metabolites in complex biologic matrixes are one of the major challenges in achieving comprehensive metabolomics data sets.[10,11] Usually samples are analyzed in a way that the concentration of the highest abundant analytes is adjusted to just reach the upper detection limit of the analytical system. Consequently, many low-abundance analytes do not reach the lower limit of detection at this concentration. We here propose automatic peak reconstruction of IVEDS from deliberately overloaded samples as a possibility to measure low abundance compounds without sacrificing the result of major analytes. To evaluate the basic response parameters of our analytical system with respect to metabolic signal quantification, we first prepared a complex standard mixture of 62 analytes, measured a dilution series of these samples and determined detection limit (LOD), quantification limit (LOQ), limit of linearity (LOL), and linear range (LR) for all metabolites (Table S1, Figure 2). The LOQ varied between 1.4 nM and 9.7 $\mu$M, reflecting differences in ionization efficiency of chemically diverse compounds. Focusing on the dynamic range covered by our analytical system, we found LR to comprise on average 2.4 orders of magnitude, ranging from 0.2 (tetradecanoic acid) to 4.2 (lysine) in a metabolite specific manner (Figure 2).

We next aimed to increase the limit of linearity by automatic correction of the raw data prior to peak identification and quantification. To this end we developed a software package reading individual data files, searching for IVEDS and modifying putatively incorrect data values. Quality control plots (QC plots) are generated for each modification and stored conveniently in a single PDF per sample. As an example, we show the main peak of Adenosine (4TMS) from the sixth dilution step with the respective $m/z$ of 556.263 Da for the M+H ion and eluting at approximately 1023 s (Figure 3). At our given conditions (scan rate, mass range, digitizer resolution) the maximum quantifiable intensity value (DS) is 971 775 counts. The software identifies all chromatographic regions in a sample where the base peak approaches this value and in the following extracts individual base peak chromatograms (BPC) for all ions within this region approaching DS as well as their first two isotopes (M+1 and M+2). In Figure 3A, these intensity values are represented by black, red, and green symbols for M, M+1, and M+2, respectively. To correct the values approaching DS (gray circles), the software determines a robust estimation of the isotopic ratios (IR) between the three BPCs using the front of the peak (scans prior to DS) and multiplies valid intensities (here the values of M+2, green color) with this IR, thereby reconstructing a peak shape similar to M+2 for the now corrected M and M+1. The result, which would be exported back into the original data files, is shown by colored lines, indicating that M+2 remains unmodified while values of M and M+1, both approaching DS, are corrected. This method works independent of the peak shape but is affected by the quality of the IR determination and by ion suppression effects on M+2. Using IR to estimate the true intensity for adenosine (4TMS) results in a peak height of $3.88 \times 10^6$.

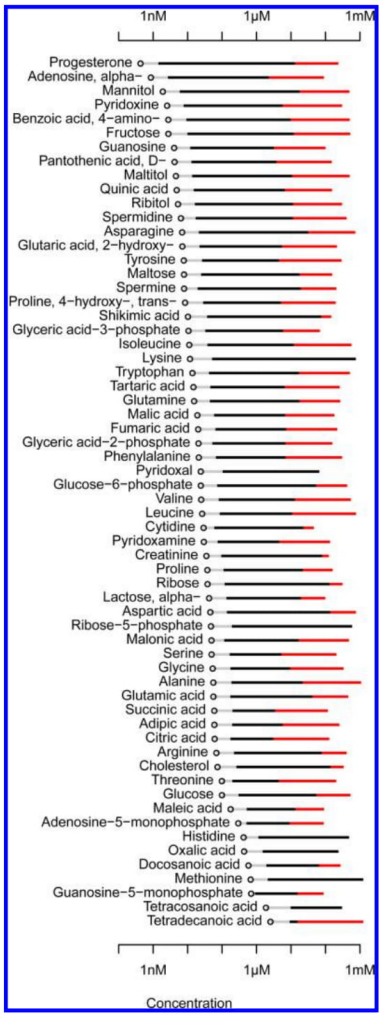



**Figure 2.** Metabolites in a complex artificial mix were quantified according to peak intensity in dense dilution series covering a range from 1.25 mM to 12.5 pM. The linear range (LR) for each metabolite, defined as the concentration range from LOQ to LOL is depicted by a black line. The LOD is indicated by a black circle. Red lines indicate an observed increase in LR for Gauss corrected raw data.



DOI: 10.1021/acs.analchem.6b02515
*Anal. Chem.* 2016, 88, 7487−7492

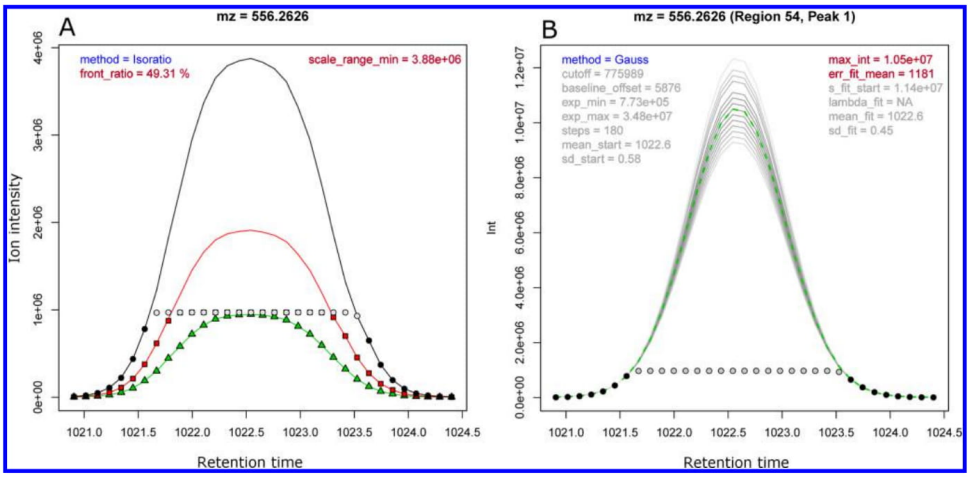



**Figure 3.** QC plots of a reconstructed peak (Adenosine 4TMS from dilution S6). Black, red, and green data points represent measurement values of ion intensity for protonated M, M+1, and M+2, respectively, while gray values indicate data points modified by the correction approach. (A) IR approach, multiplying the values of an isotopic peak (here M+2, green color) by a factor determined from the ratio of M+2 and M in the peak front to reconstruct the M peak. (B) Gauss approach, fitting a Gaussian curve to reconstruct M by minimizing the mean error to the valid measurement data points of M.

As an alternative, we utilized the fact that most IVEDS in GC-APCI-MS are not yet chromatographically overloaded and hence maintain a Gaussian peak shape at the peak basis. This allows one to fit the parameters of a Gaussian curve (mean $m$, standard deviation sd, and height $h$) to the observed data, by minimizing the residual error and starting from different initial values. The result of such an automatic Gauss fit for adenosine (4TMS) is shown in Figure 3B. Analogous to Figure 3A, the measured raw data are depicted by black symbols, and the identified IVEDS are emphasized in gray, while the reconstructed peak shape is shown by a green dashed line. Thin gray lines indicate alternative solutions of a Gaussian fit with higher residual error. The two parameters of highest importance annotating the plot are the mean error of the fit and the maximum intensity (upper right corner). The mean error is calculated from the deviations of the observed data points from the values we assigned based on the optimal Gaussian curve, excluding the ones approaching DS. Values in the range of 0.1 to 1% of DS (here around $10^3$) are usually very good. After finding the optimal parameters by minimizing the mean error, we substitute all values approaching DS with their calculated values, respectively. In the example this leads to a reconstructed peak maximum of $1.05 \times 10^7$, being about 3-fold higher as the solution obtained by the IR based method and nearly 10 times higher than DS. QC plot files for all samples are provided as Supporting Information.

Crucial for a correct reconstruction of analytical signals are sensible starting parameters including a sufficient number of data points in peak front or tail as well as a good shape of the measured peak basis. GC-APCI-MS is advantageous in that respect, as detector saturation is usually reached much earlier than chromatographic saturation. Thus, for most peaks, we observed nearly Gaussian shape even when signals were exceeding DS more than 10-fold. Only in cases of decreased chromatographic performance, e.g. when using splitless injection of concentrated samples, we found more peaks deviating from a standard Gaussian, which were better to reconstruct using an exponentially modified Gaussian (EMG), an option available in the software. However, in splitless injection we observed several peaks with a delayed intensity decrease in peak tail, leading to extreme parameters of the EMG and larger reconstruction errors. Moderate examples of distorted peak shapes are shown in Figure S2A,B. Especially in metabolite dense chromatographic areas peak shapes can be found distorted, due to interference of coeluting signals. The software can partially cope with that issue by detecting skewed front or tail values and limiting the Gaussian fit to one side of the peak only (Figure S2C). A systematic evaluation of the capabilities of Gauss based reconstruction indicates that the log-ratio $R$ between data at the peak basis used for reconstruction and reconstructed data should not be lower than 2, i.e., a peak with baseline at $10^3$ in an analytical system where DS = $10^6$ can be reconstructed with moderate error if overloaded 50-fold ($R = 6 - 3/7.5 - 6 = 3/1.5 = 2$) but not if overloaded 100-fold ($R = 3/2$) or higher (Data S1).

Processing all metabolite peaks of the complete dilution series in the above-described fashion allows one to compare systematically LOD, LOQ, and LOL in originally measured data as well as in data which was modified by the IR method or the Gauss approach (Table S1). For adenosine (4TMS), the result is shown in Figure 4, identifying a LOQ of 2.68 nM and a LOL of 2.28 $\mu$M for the original data, leading to a linear range (LR) of 2.93. IR or Gauss correction modifies IVEDS and, thereby, extends the detectable LR to 3.65 and 4.51, respectively, which is equivalent to an absolute and relative increase of 1.58 orders of magnitude or 54% for Gauss corrected data.

Detailed per metabolite evaluations are given in File S2. Over all metabolites, we determine a median increase in LR of 1.4 orders of magnitude comparing Gauss corrected and raw data,



DOI: 10.1021/acs.analchem.6b02515
Anal. Chem. 2016, 88, 7487−7492

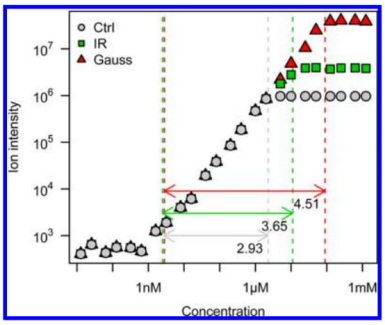



**Figure 4.** Comparison of peak intensities for $m/z$ = 556.2625 (which represents the M+H peak of the 4TMS derivate of adenosine) from samples of a dilution series with 24 steps. The linear range was determined as the difference between LOQ and LOL in log-scale and is indicated in color corresponding to the applied sample processing (gray, raw data; green, IR approach; red, Gauss approach).

evaluating those peaks where the target ion approached DS in at least one sample and therefore was subjected to the algorithm (Figure 5A). For the IR approach, the average gain was more moderate (average LR increase of 0.6). Figure 5B shows the relation between the observed gain in LR and the maximum gain which could have theoretically been achieved within this experiment. For two-thirds of metabolites we realize more than 50% of the potential gain, while 10 metabolites perform poorly with less than 25% realized gain. The performance is independent of chemical class (represented by different symbols in the plot) and the absolute gain approaches a limit at about 2 orders of magnitude.

To assess the performance of the algorithm systematically in a more applied example, we prepared a dilution series of a blood serum sample where we adjusted the processed volume of the highest concentration to contain several metabolites with IVEDS. We processed the raw data of all 7 samples of the dilution series using CorrectOverloadedPeaks and standard parameters. We next investigated all ion signals which were modified (IVEDS) in the highest concentration sample, if they were of biological origin, i.e., they showed lower intensity values in lower concentrations in contrast to contaminating peaks from chemical reagents which are present more or less constant over all samples. After fitting robust linear models to these calibration curves, we calculated the residual error for IVEDS of in total 38 separate compounds which comprise 59 different fragments and up to 2 additional isotopes per fragment (File S3). This approach is very close to the anticipated use of CorrectOverloadedPeaks, where we allow a small number of abundant compounds in a sample to exceed the analytical detection limit to some extent (50-fold) to ultimately gain more information about numerous small abundant compounds. In our example, we determined the residual error of reconstructed peaks to be below 20% in more than 90% of all cases, confirming the results from our defined chemical mixture (Figure S4).

While the Gauss approach yielded better results in successfully correcting IVEDS up to 50-fold above DS compared to 10-fold for IR approach, using the signal intensity of isotopes below DS to reconstruct a peak is a very robust alternative to Gaussian or EMG reconstruction. Even highly

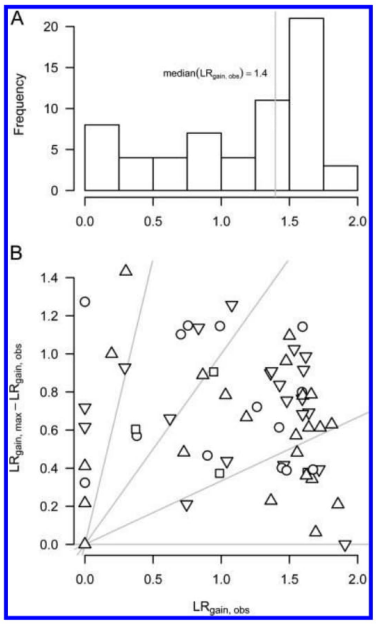



**Figure 5.** Absolute increase in LR after Gauss correction ($LR_{gain,\ obs}$) over all metabolites depicted as (A) a histogram with the median value indicated by a gray line or (B) relative to the difference between the maximum possible gain within this experiment ($LR_{gain,\ max}$), i.e., the ratio between the maximum metabolite concentration of 1.25 mM and the LOL detected in the control settings. Different symbols represent chemical classes (□, sugar; △, organic acid; ▽, amino acid; ○, others). Gray lines indicate the relative amount of the maximum possible gain which was achieved applying Gauss correction, i.e., for the 14 data points below the 75% line Gauss correction extended LR by more than 75% of what was theoretically possible, given that our maximum concentration was limited at 1.25 mM.

skewed peak shapes can be reconstructed using this method, rendering it preferable for chromatographically poor separations, e.g., as encountered in some LC−MS applications. The IR approach should also be used in flux analysis, where it is essential that isotope ratios are maintained during correction because they are used to calculate enrichment. In TOF analyzers, the isotopic ratio is considered stable[12] with an average error of less than 2%. However, we could show that isotopic ratio has a negative bias in peak tails so scan rates yielding at least three data points within the peak front (>5 Hz in our case) are required to get a robust estimate of isotopic ratio, and Orbitrap technology may not be suited to this approach as isotopic stability is considered less good there.[12]

Kalambet et al. investigated peak reconstruction for LC−MS using differently parametrized EMG function and observed for a manually selected set of 44 peaks which were artificially limited to 10% of their original size an average reconstruction error of 20% for peak area, which is similar to our observations.[6] However, they applied their results only to a small dilution series of one individual analyte (Nipagin, highest concentration 5-fold above detector limit), showing that an a priori known peak shape allows for lowest reconstruction errors


DOI: 10.1021/acs.analchem.6b02515
Anal. Chem. 2016, 88, 7487−7492

in comparison to less constrained models, which is in agreement with our findings regarding the quality of starting parameters being crucial for the Gauss approach.

Alternative strategies to peak reconstruction comprise dilution of each sample, either prior to derivatization or using a split ratio in GC, and quantification on fragment or isotopic masses,[13] which are well below the detector limit for all samples within the experiment. However, these methods are labor intensive and time-consuming, while the solution presented here can be included in an existent data analysis pipeline as an automatic preprocessing step without additional manual work or further modifications. Alternative strategies are not free of error which we partially recapitulate with our IR approach and investigated for the application of different split ratios (Figure S3).

An overloaded signal may be the result of various steps within the analytical chain and among others occur due to problems in chromatography, ionization, and ion detection.[14] Chromatographic overloading, indicated by a steep linear peak front, is expected to be a rare event in GC-APCI-MS as high detector sensitivities will usually require only small on-column amounts. Reconstruction of a peak from such a linear peak front would be impaired and can only partly be encounted for by removing information from the peak front. Ion suppression occurs constantly in soft ionization techniques (APCI, ESI) as the ion source has only a limited capacity to generate charged molecules.[15] In consequence, large numbers of molecules eluting from the separation column will compete for the available ionization energy and show reduced intensities. This process is complex and modeling it is outside the scope of this manuscript. While it hampers a quantitative peak reconstruction, its influence on individual samples can be expected to be similar, thus, still allowing relative quantification after reconstruction. Finally, electron multiplier and digitizer used as detector and converter in TOF instruments have a limited capacity to count and sum up individual TOF events, leading to flat-topped peaks. Here, we showed that this limitation can be well overcome computationally using stable isotopic ratios or Gaussian modeling.

## ■ CONCLUSIONS

Peak reconstruction will not be the method of choice in experiments where absolute quantitative data are required. However, for all metabolic screening experiments where relative quantification is used to identify potential biomarkers, peak reconstruction will allow one to drastically increase throughput, dynamic concentration range covered and, thereby, total number of metabolites analyzed without relevant additional costs, time, or complicated data integration steps at an error rate well within the boundaries of machine performance.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b02515.

Figures S1–S4, overview Table S1 presenting LOD/LOQ/LOL data for all metabolites in number as a complementary information to Figure 2, additional detailed methods regarding sample preparation and GC/MS conditions, and a systematic evaluation of the Gauss algorithm on artificial data (PDF)

PDF QC plot data for all samples of the dilution series as produced by the software (ZIP)

Detailed LOD/LOQ/LOL analyses for all metabolites investigated in this study (PDF)

Detailed investigations for various analytes from a dilution series of a biological sample (blood serum) (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: jan.lisec@charite.de. Fax: 0049 (30) 450559975.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank the Berlin School of Integrative Oncology (BSIO) within the German Excellence Initiative for funding to Friederike Hoffmann.

## ■ REFERENCES

(1) Beecher, C. W. W., Harrigan, G. G., Goodacre, R., Eds.; Springer US: Boston, MA, 2003; pp 311–319.
(2) Madsen, R.; Lundstedt, T.; Trygg, J. *Anal. Chim. Acta* **2010**, *659*, 23–33.
(3) Mamas, M.; Dunn, W. B.; Neyses, L.; Goodacre, R. *Arch. Toxicol.* **2011**, *85*, 5–17.
(4) Dunn, W. B.; Bailey, N. J. C.; Johnson, H. E. *Analyst* **2005**, *130*, 606–625.
(5) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801–D807.
(6) Kalambet, Y.; Kozmin, Y.; Mikhailova, K.; Nagaev, I.; Tikhonov, P. *J. Chemom.* **2011**, *25*, 352–356.
(7) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.
(8) Konieczka, P.; Namieśnik, J. *Quality Assurance and Quality Control in the Analytical Chemical Laboratory: A Practical Approach*; CRC Press: Boca Raton, FL, 2009.
(9) Sonderegger, D., 2012, http://CRAN.R-project.org/package=SiZer.
(10) Wang, Y.; Liu, S.; Hu, Y.; Li, P.; Wan, J.-B. *RSC Adv.* **2015**, *5*, 78728–78737.
(11) Mirnaghi, F. S.; Caudy, A. A. *Bioanalysis* **2014**, *6*, 3393–3416.
(12) Glauser, G.; Veyrat, N.; Rochat, B.; Wolfender, J.-L.; Turlings, T. C. J. *J. Chromatogr. A* **2013**, *1292*, 151–159.
(13) Wang, Y.; Hu, H.; Su, Y.; Zhang, F.; Guo, Y. *J. Sep. Sci.* **2016**, *39*, 1137.
(14) Jaulmes, A.; Ignatiadis, I.; Cardot, P.; Vidal-Madjar, C. *J. Chromatogr. A* **1987**, *395*, 291–306.
(15) Remane, D.; Meyer, M. R.; Wissenbach, D. K.; Maurer, H. H. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 3103–3108.



DOI: 10.1021/acs.analchem.6b02515
*Anal. Chem.* 2016, 88, 7487–7492

analytical chemistry

Technical Note

pubs.acs.org/ac

# Automated Annotation and Evaluation of In-Source Mass Spectra in GC/Atmospheric Pressure Chemical Ionization-MS-Based Metabolomics

Carsten Jaeger,[*,†,‡] Friederike Hoffmann,[†] Clemens A. Schmitt,[†,‡,§] and Jan Lisec[†,∥]

[†]Charité - Universitätsmedizin Berlin, Medical Department of Hematology, Oncology, and Tumor Immunology, and Molekulares Krebsforschungszentrum (MKFZ), Augustenburger Platz 1, 13353 Berlin, Germany

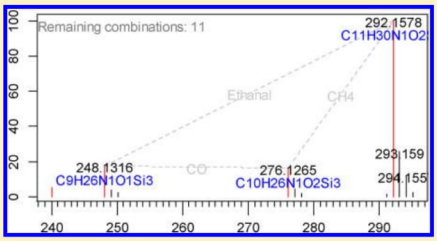
[‡]Berlin Institute of Health (BIH), Kapelle-Ufer 2, 10117 Berlin, Germany

[§]Max-Delbrück-Center for Molecular Medicine (MDC), Robert-Rössle-Straße 10, 13125 Berlin, Germany

[∥]German Cancer Consortium, Deutsches Krebsforschungszentrum (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

Supporting Information

**ABSTRACT:** Gas chromatography using atmospheric pressure chemical ionization coupled to mass spectrometry (GC/APCI-MS) is an emerging metabolomics platform, providing much-enhanced capabilities for structural mass spectrometry as compared to traditional electron ionization (EI)-based techniques. To exploit the potential of GC/APCI-MS for more comprehensive metabolite annotation, a major bottleneck in metabolomics, we here present the novel R-based tool *InterpretMSSpectrum* assisting in the common task of annotating and evaluating in-source mass spectra as obtained from typical full-scan experiments. After passing a list of mass-intensity pairs, *InterpretMSSpectrum* locates the molecular ion ($M_0$), fragment, and adduct peaks, calculates their most likely sum formula combination, and graphically summarizes results as an annotated mass spectrum. Using (modifiable) filter rules for the commonly used methoximated-trimethylsilylated (MeOx-TMS) derivatives, covering elemental composition, typical substructures, neutral losses, and adducts, *InterpretMSSpectrum* significantly reduces the number of sum formula candidates, minimizing manual effort for postprocessing candidate lists. We demonstrate the utility of *InterpretMSSpectrum* for 86 in-source spectra of derivatized standard compounds, in which rank-1 sum formula assignments were achieved in 84% of the cases, compared to only 63% when using mass and isotope information on the $M_0$ alone. We further use, for the first time, automated annotation to evaluate the purity of pseudospectra generated by different metabolomics preprocessing tools, showing that automated annotation can serve as an integrative quality measure for peak picking/deconvolution methods. As an R package, *InterpretMSSpectrum* integrates flexibly into existing metabolomics pipelines and is freely available from CRAN (https://cran.r-project.org/).




Gas chromatography coupled to mass spectrometry using electron ionization (GC/EI-MS) is widely used in metabolomics, as it combines efficient chromatographic separation with straightforward metabolite annotation based on comprehensive mass spectral libraries.[1−3] Despite continuous improvement of preprocessing tools and spectral libraries, however, many analytes in nontargeted experiments remain unidentified.[4] A new generation of recently introduced atmospheric pressure chemical ionization (APCI) interfaces promises significant advances for this problem of commonly observed "unknown" peaks. GC/APCI mass spectra, in contrast to EI spectra, contain dominating molecular ions ($[M + H]^+$) that often allow *de novo* identification of metabolites through sum formula prediction and structural MS/MS experiments.[5,6] In terms of overall metabolite coverage, GC/APCI-MS proved equivalent to GC/EI-MS and even largely surpassed the latter in terms of sensitivity.[7] Consequently, use of GC/APCI-MS in metabolomics has increased strongly in recent years.[8,9]

Metabolomics workflows usually comprise a number of computational steps following data acquisition, including peak picking, peak alignment, chromatographic deconvolution, and compound annotation. The software tools used for these steps have mostly been developed and optimized for either GC with hard ionization (GC/EI-MS) or liquid chromatography−mass spectrometry using soft ionization (LC/ESI-MS). Specific solutions for GC/APCI-MS are just only becoming available. A first MS/MS library for GC/APCI-MS, for example, has recently been introduced.[10] Such libraries allow annotation of measured spectra based on precursor and qualifying ions

Received: July 18, 2016
Accepted: September 1, 2016
Published: September 1, 2016





DOI: 10.1021/acs.analchem.6b02743
Anal. Chem. 2016, 88, 9386−9390

information similar to commonly employed LC/ESI-MS strategies, e.g., using Metlin. Currently, however, the Leiden library contains only relatively few compounds (138 with $MS^1$, 106 with $MS^2$ spectra as of February 2016). To circumvent the issue of missing reference spectra, Ruttkies et al.[11] presented a workflow based on *in silico* derivatization and fragmentation of all KEGG (or optionally Pubchem) compounds. This strategy allowed correct assignment of metabolite structures for 57% of the 104 tested metabolites. Still, manual inspection and interpretation of GC/APCI-MS spectra remains a frequent task for analytes without available information in databases or where experimental spectra deviate from reference ones. Software tools that automate the time-consuming tasks of precursor/fragment assignment, sum formula prediction, and similar annotation, however, are currently lacking for GC/APCI-MS.

In the present paper, we introduce *InterpretMSSpectrum*, a novel R package that automates spectral annotation for high-resolution GC/APCI mass spectra. Within a mass-intensity list, *InterpretMSSpectrum* locates the molecular and fragment peaks and establishes likely neutral loss/adduct interrelationships based on well-established chemical rules. These are further used to limit the otherwise large number of sum formula suggestions, and results are returned to the user in a concise graphical summary. We demonstrate the utility of *InterpretMSSpectrum* by annotating 86 in-source mass spectra of frequently encountered primary metabolites, yielding correct (rank-1) sum formula prediction in 84% of the cases. Furthermore, we use *InterpretMSSpectrum* to evaluate the purity of pseudospectra generated by four popular metabolomics preprocessing tools, showing that automated annotation can serve as an integrative quality measure for peak picking/deconvolution pipelines.

## ■ EXPERIMENTAL SECTION

**Sample Preparation, Derivatization, and GC/APCI-MS Analysis.** A standard mixture of 59 metabolites (see Table S-1) was prepared at concentrations of 78, 19.5, 4.9, 1.2, and 0.3 $\mu$M, respectively, derivatized by methoximation/trimethylsilylation and analyzed by GC/APCI-MS as described in detail in Methods S-1.

**Data Analysis.** Data files (Bruker .d) were recalibrated in DataAnalysis 4.2 (Bruker Daltonik GmbH, Germany) against siloxane background ions ($m/z$ 207.0324, 223.0637, 347.0950, 445.1200, 519.1576, 593.1576) and exported as netCDF exchange format. For peak detection and chromatographic deconvolution, xcms/CAMERA[12,13] was applied as detailed in Table S-2. Resulting pseudospectra were identified on the basis of $m/z$ (±4 ppm) and retention time (±5 s). Of the five concentrations analyzed, the spectrum with maximum base peak intensity within $1 \times 10^4$ to $8 \times 10^5$ counts, corresponding to 1−80% detector saturation, was selected as a reference spectrum for further analysis with *InterpretMSSpectrum*.

**Automated Annotation Approach.** An algorithm for automated mass spectral annotation was implemented as an R (https://www.r-project.org/) function, accepting an arbitrary list of mass-intensity pairs (spectrum) as input. Possible sum formulas for each informative peak (peaks above a certain threshold) are calculated *de novo*, applying different elemental filters empirically derived from all entries of the Golm Metabolome Database (GMD).[3] Remaining formulas are filtered pairwise according to typical neutral losses of methoximated/silylated compounds (or other user-supplied rulesets), and final potential formula combinations are scored integratively based on mean mass deviation of potential $[M + H]^+$ and fragments. Results for the highest scoring combination are summarized graphically in the form of an annotated spectrum. Details on the algorithm and use of the package are described in Methods S-1.

**Comparison of Different Deconvolution Algorithms.** To compare spectral deconvolution by different software tools, the above standard mixture was spiked into a methanolic extract of human plasma and analyzed as above. The resulting netCDF file was processed with DataAnalysis 4.2 (Bruker Daltonik GmbH, Germany) using the "Dissect" algorithm, Maven (Build 682),[14] MET-COFEA (Beta 2013-10-15)[15] and xcms (1.46)/CAMERA (1.26),[12,13] and pseudospectra were exported for further use. Processing parameters in each of these packages were set according to official recommendations or personal communication with the authors (Table S-2).

## ■ RESULTS AND DISCUSSION

**Automated Annotation of GC/APCI-MS Spectra.** GC/APCI-MS yields intense molecular ion peaks for most compound classes including the commonly used trimethylsilyl derivatives, improving structural elucidation of unknown metabolites as compared to conventional GC/EI-MS analysis. Typically, silylated compounds also exhibit moderate formation of in-source fragments under APCI conditions[6] that can contribute important structural clues for annotation. The GC/APCI-MS spectrum of glucose (1MeOx) (5TMS), for example, contains 14 major ions including the molecular peak at $m/z$ 570.2949 and dominant fragments at $m/z$ 554.2636, $m/z$ 480.2448, $m/z$ 390.1949, $m/z$ 319.1578, and $m/z$ 307.1579 (Figure 1). Manual interpretation of the spectrum

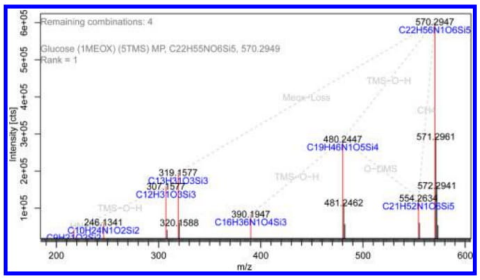



**Figure 1.** GC/APCI-MS spectrum of glucose 1MeOx 5TMS ($C_{22}H_{55}NO_6Si_5$, $m/z$ 570.2949) annotated by *InterpretMSSpectrum*. Highlighted features include major mass peaks (red), predicted sum formulas (blue), and calculated neutral loss relationships (gray). "Remaining combinations" designates the final number of possible sum formula combinations of the major mass peaks after filtering for elemental composition and assumed neutral loss relationships of the fragments. The correct formula, passed as an optional argument to *InterpretMSSpectrum*, was found on rank 1 among these candidates.

reveals some of these fragments as typical neutral losses for silylated compounds under CI conditions,[16] e.g., $[M+H-CH_4]^+$ ($m/z$ 554.2636), $[M+H-TMSOH]^+$ ($m/z$ 480.2448), and $[M+H-(TMSOH)_2]^+$ ($m/z$ 390.1949). *InterpretMSSpectrum* recapitulates such manual annotation in an automated way and uses the detected peak relationships for sum formula determination (for a scheme of the algorithm, see Figure S-1). In addition to the losses mentioned, *InterpretMSSpectrum*


DOI: 10.1021/acs.analchem.6b02743
*Anal. Chem.* 2016, 88, 9386−9390

annotated a characteristic methoxyamine (Meox) loss ([M+H−$C_9H_{25}NO_3Si_2$]$^+$) resulting from the methoximated aldehyde group of the sugar, as well as several secondary losses of $CH_4$, O−DMS, and TMS−OH groups. The more plausible chemical relationships can be established, the more sum formula candidates can be excluded on the basis of mutual comparisons: from 288 candidate formulas before neutral loss filtering, only 4 suggestions were kept (1.3%). In case none of the defined rules applies to a given pair of peaks, all calculated formulas will be considered, avoiding discrimination of less common fragmentation paths. *InterpretMSSpectrum* displays the final number of candidates in the annotated spectrum together with the rank of the correct sum formula among these suggestions. In the example, the correct sum formula ($C_{22}H_{55}NO_6Si_5$) was found on rank 1, with an integrative score significantly higher than the lower-ranked candidates (94.1 vs 87.5, 78.0, and 74.1, respectively).

We tested automated annotation for 86 GC/APCI spectra of 59 primary metabolites acquired on a high-resolution quadrupole time-of-flight (HR-QTOF) system. Using moderate ion source parameters, we obtained in-source spectra with 1−27 major (deisotoped) peaks (median: 5), a majority of which could be explained in terms of 22 predefined neutral losses (Table S-3). For example, $CH_4$ losses occurred in 74% of the spectra and TMS−OH and O−DMS losses in 52% and 45% of the spectra, respectively (Figure 2A). As expected, the number of initial sum formula suggestions for molecular ion ($M_0$) increased exponentially with the molecular mass, ranging between 33 for $m/z$ 188.1102 and 52870 for $m/z$ 948.4643 (Figure 2B). Filtering by elemental composition reduced these by 69% to 97% (median 89%), and additionally filtering by neutral losses by 69% to 99% (median 98%). The number of remaining candidates ranged between 1 for small metabolites such as alanine (2TMS) and 884 for the relatively large guanosine-5-monophosphate (6TMS) ($M_r$ = 795.2952), among which the correct formula was found on rank 1 for 72 of the 86 spectra (84%) and on ranks 2−3 for another 7% (Figure 2C). By comparison, when *InterpretMSSpectrum* was applied to the molecular ion and its isotopes alone, only 63% and 19% of the suggestions were found on ranks 1 and 2−3, respectively, demonstrating the informative value of in-source fragments. Established peak relationships for the full set of spectra are provided in Data S-1.

A number of tools have implemented computational annotation of mass spectra with different use cases in mind. MS2Analyzer,[17] for example, searches many MS/MS spectra for user-defined mass differences (neutral losses, precursor/product ion transitions) to detect characteristic features of particular compound classes. mzGroupAnalyzer[18] detects possible metabolic steps related to chemical and biochemical transformations in the $m/z$ features of an experiment and annotates putative substructures in the respective mass spectra. CAMERA[12] extracts pure compound spectra, annotates isotope and adduct ions according to user-defined lists, and generates mass hypotheses of the underlying compound. Despite some overlap in functionalities with our approach, none of the packages offers automated rule-constrained generation of sum formula combinations for precursor and fragment ions and out-of-the-box compatibility with methoximated-trimethylsilylated (MeOx-TMS) derivatives. MS/MS filtered sum formula generation is implemented in the SmartFormula3D algorithm (Bruker Daltonik GmbH, Germany) as well as in MZmine 2.[19] However, none of these use customizable mass difference lists

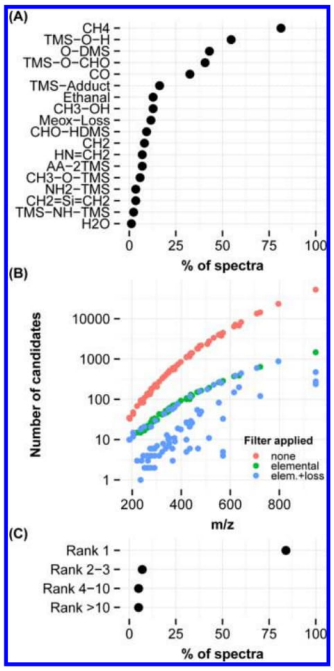



**Figure 2.** Summary of annotation results for 86 GC/APCI-MS spectra of standard compounds. (A) Frequency of neutral losses detected within the spectra. (B) Number of sum formula candidates as a function of mass-to-charge ratio, obtained before and after filtering. Two filters were applied sequentially: (1) an elemental composition filter and (2) a filter for typical neutral losses of silylated compounds. (C) Rank distribution of the correct sum formulas within the candidate lists of the respective compounds.

(neutral losses, adducts) to filter by chemically "plausible" candidates. In addition, they lack the graphical summary capabilities of *InterpretMSSpectrum* that facilitate interpretation "at a glance" of results by the user. Being conceptually different, we do not compare our tool to *in silico* fragmentation algorithms such as SIRIUS[20] or MetFrag,[21] which potentially can also differentiate structural isomers, which none of the other tools including ours is capable.

**Comparison of Different Deconvolution Algorithms.** As automated annotation scores chemically plausible fragments, we next investigated how contaminated spectra influence results. Different sets of pseudospectra were obtained by again analyzing a mixture of primary metabolite standards, but this time spiked into a biological matrix (human blood plasma) to increase chromatographic complexity. Three different open-source (Maven, MET-COFEA, xcms/CAMERA) and one commercial algorithm (Dissect) were used to detect features and deconvolute compounds in this chromatogram. As shown in Figure 3A, algorithms differed noticeably in the reduction of sum formula candidates achieved by filtering, with xcms/CAMERA spectra yielding on average the smallest number of final sum formula candidates of all packages. This indicated that spectra contained different numbers of structurally informative


DOI: 10.1021/acs.analchem.6b02743
*Anal. Chem.* 2016, 88, 9386−9390

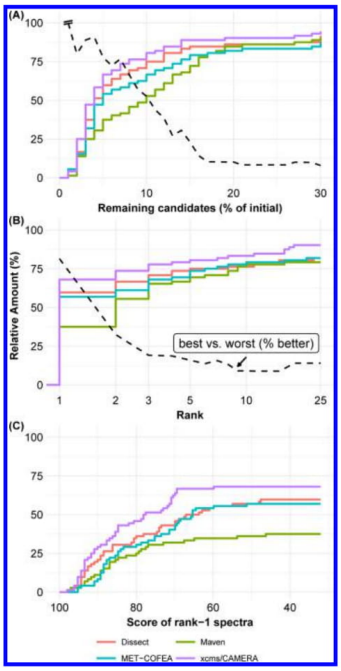



**Figure 3.** Evaluation of pseudospectra by automated annotation. Metabolite standards were spiked into human plasma, and pseudospectra resulting from the application of the tools indicated were evaluated with InterpretMSSpectrum ($n$ = 72; 100%). (A) Remaining sum formulas after the filtering steps of *InterpretMSSpectrum* as a percentage of initial sum formulas. (B) Ranks of correct sum formula within the compounds' candidate lists. (C) Mean score of sum formula combinations for correctly predicted (rank-1) spectra. Values in (A) to (C) are represented as cumulative frequencies; broken lines in (A) and (B) additionally indicate how much better the best algorithm scored compared to the worst.

peaks, allowing one to exclude false sum formula hypotheses to different degrees. The number of correct sum formula annotations, i.e., the cases where the correct sum formula was found on rank 1 within the candidate list, also reflected this (Figure 3B), with xcms/CAMERA yielding correct annotations for 71% of the spectra, in contrast to only 40−65% for the other algorithms. The score of these rank-1 spectra, indicating the mean agreement of measured with hypothetical masses/isotope patterns, was also highest for the xcms/CAMERA pipeline (Figure 3C), suggesting that mass and isotope pattern information were more precisely extracted than by the remaining tools.

Our results suggest that automated annotation can serve as a quality measure for metabolomic data preprocessing as it integrates the quality of peak picking, chromatographic deconvolution, and possibly other steps involved in preprocessing. Such evaluation is a nontrivial but highly important task to avoid bias in high-throughput data. Zhang and Zhao[22] compared different measures for peak quality and found that their "zigzag index" performed best for discriminating good from bad peaks. Libiseller et al.[23] developed the R package "IPO" that evaluates peak picking results by the completeness of the expected isotopic peaks. While these approaches focus on individual mass peaks, *InterpretMSSpectrum* scores entire mass spectra based on chemically plausible rules. Both missing (false negative) and contaminating (false positive) mass peaks worsen the score of a pseudospectrum. For Aspartic acid (3TMS), for example, the absence of $m/z$ 243.0961 in the spectrum from Maven led to more remaining candidates and worse rank score as compared to the spectra of the other tools (Figure S-2). Similarly, contaminating peaks in the Ribose (1MeOx) (4TMS) spectrum from Dissect resulted in inferior scores compared to the other tools (Figure S-3). Such spectral comparisons can likewise be carried out for different analytical conditions such as chromatographic or ion source settings, making *InterpretMSSpectrum* useful for general method optimization.

Our results also demonstrate that peak detection and deconvolution tools are differently suited for a novel analytical technique such as GC/APCI-MS. An algorithm optimized for LC/ESI-MS chromatograms might perform worse for GC/APCI-MS, as peak widths fall into the lower, nonoptimized end of the configurable range. While the typical Gaussian GC peak should create little problems to peak finders, chromatographic background, e.g., from column bleed might be an issue. Such background is more pronounced in GC/APCI-MS than in GC/EI-MS due to higher overall sensitivity of the method.[7] We consequently observed that some peak pickers, e.g., the centWave algorithm[24] used by xcms, had a high false positive rate in the background-intense region during the column bake-out (Figure S-4), which was likely due to the particular combination of narrow peak widths and noise frequencies. Other issues were related to the different isotope spacing (mass distances <1 amu) of silylated compounds as compared to nonderivatized metabolites containing only C, H, N, O, P, and S, as deconvolution algorithms are frequently optimized for mass distances >1 amu for better specificity. Some pseudospectra, for example, often lacked the M + 1 or M + 2 peaks, suggesting partial discrimination of Si-containing isotopologues. This further highlights the importance of optimization studies prior to analyzing large-scale experiments.

## ■ CONCLUSIONS

We here present the novel R package *InterpretMSSpectrum* that aims to fill a gap in software tools supporting systematic analysis of GC/APCI-MS metabolomics data sets. *InterpretMSSpectrum* automates common manual steps in the annotation of mass spectra: selection of relevant peaks, establishment of structural relationships, and calculation of sum formulas. We envisage several use cases for *InterpretMSSpectrum*. First, as demonstrated for a set of standard compounds, it assists in structural annotation of fragments and adducts, a recurring task in fragmentation studies or buildup of annotated libraries. Second, it supports sum formula calculation in consideration of elemental and structural constraints, which facilitates reviewing candidate compounds, e.g., from conventional EI libraries such as NIST or GMD. Given the high true positive rate of 84% in correct sum formula prediction, we are confident that a similar rate is achieved in unknown analysis. Third, it allows comparing sets of spectra acquired under different analytical conditions or, as demonstrated here, obtained from different preprocessing tools, which facilitates optimization of the analytical and data analysis parts of metabolomics protocols. *InterpretMSSpectrum* is used here


DOI: 10.1021/acs.analchem.6b02743
*Anal. Chem.* 2016, 88, 9386−9390

for TMS in-source spectra from a high-resolution QTOF mass spectrometer, and we expect it to be likewise applicable to MS or MS/MS spectra from other high-resolution instruments (TOF, Orbitrap) and to spectra of different compound classes (e.g., PAHs, PCBs) once appropriate neutral loss tables are supplied. As an R package, it flexibly integrates into existing analysis pipelines.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b02743.

Figures S-1 to S-4, Tables S-1 to S-3, detailed Materials and Methods (Methods S-1), and the full set of annotated reference spectra (Data S-1) (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

*Phone: +49(30)450559106. E-mail: carsten.jaeger@charite.de.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank Steffen Neumann (IPB Halle, Germany) for fruitful discussions about current algorithmic approaches in metabolite annotation and for critical comments on an early version of this manuscript. We thank the Berlin School of Integrative Oncology (BSIO) within the German Excellence Initiative for funding to Friederike Hoffmann.

## ■ REFERENCES

(1) Fiehn, O. *Plant Mol. Biol.* **2002**, *48* (1/2), 155−171.
(2) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. *Anal. Chem.* **2009**, *81* (24), 10038−10048.
(3) Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmuller, E.; Dormann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D. *Bioinformatics* **2005**, *21* (8), 1635−1638.
(4) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, *9* (S1), 44−66.
(5) Carrasco-Pancorbo, A.; Nevedomskaya, E.; Arthen-Engeland, T.; Zey, T.; Zurek, G.; Baessmann, C.; Deelder, A. M.; Mayboroda, O. *A. Anal. Chem.* **2009**, *81* (24), 10071−10079.
(6) Strehmel, N.; Kopka, J.; Scheel, D.; Böttcher, C. *Metabolomics* **2014**, *10* (2), 324−336.
(7) Wachsmuth, C. J.; Almstetter, M. F.; Waldhier, M. C.; Gruber, M. A.; Nürnberger, N.; Oefner, P. J.; Dettmer, K. *Anal. Chem.* **2011**, *83* (19), 7514−7522.
(8) Li, D.-X.; Gan, L.; Bronja, A.; Schmitz, O. *J. Anal. Chim. Acta* **2015**, *891*, 43−61.
(9) Lisec, J.; Hoffmann, F.; Schmitt, C.; Jaeger, C. *Anal. Chem.* **2016**, *88* (15), 7487−7492.
(10) Pacchiarotta, T.; Derks, R. J. E.; Hurtado-Fernandez, E.; van Bezooijen, P.; Henneman, A.; Schiewek, R.; Fernández-Gutiérrez, A.; Carrasco-Pancorbo, A.; Deelder, A. M.; Mayboroda, O. A. *Bioanalysis* **2013**, *5* (12), 1515−1525.
(11) Ruttkies, C.; Strehmel, N.; Scheel, D.; Neumann, S. *Rapid Commun. Mass Spectrom.* **2015**, *29* (16), 1521−1529.
(12) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84* (1), 283−289.
(13) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779−787.
(14) Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc: Hoboken, NJ, USA, 2002.
(15) Zhang, W.; Chang, J.; Lei, Z.; Huhman, D.; Sumner, L. W.; Zhao, P. X. *Anal. Chem.* **2014**, *86* (13), 6245−6253.
(16) Warren, C. R. *Metabolomics* **2013**, *9* (S1), 110−120.
(17) Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86* (21), 10724−10731.
(18) Doerfler, H.; Sun, X.; Wang, L.; Engelmeier, D.; Lyon, D.; Weckwerth, W. *PLoS One* **2014**, *9* (5), e96188.
(19) Pluskal, T.; Uehara, T.; Yanagida, M. *Anal. Chem.* **2012**, *84* (10), 4396−4403.
(20) Dührkop, K.; Scheubert, K.; Böcker, S. *Metabolites* **2013**, *3* (2), 506−516.
(21) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11* (1), 148.
(22) Zhang, W.; Zhao, P. X. *BMC Bioinf.* **2014**, *15* (Suppl11), S5.
(23) Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; Magnes, C. *BMC Bioinf.* **2015**, *16* (1), 736.
(24) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9* (1), 504.



DOI: 10.1021/acs.analchem.6b02743
*Anal. Chem.* 2016, 88, 9386−9390

# Nontargeted Identification of Tracer Incorporation in High-Resolution Mass Spectrometry



analytical chemistry

Article

Cite This: *Anal. Chem.* 2018, 90, 7253–7260

pubs.acs.org/ac

## Nontargeted Identification of Tracer Incorporation in High-Resolution Mass Spectrometry

Friederike Hoffmann,[†] Carsten Jaeger,[†,‡] Animesh Bhattacharya,[†] Clemens A. Schmitt,[†,‡,§] and Jan Lisec*[,∥]

[†]Charité-Universitätsmedizin Berlin, Medical Department of Hematology, Oncology, and Tumor Immunology and Molekulares Krebsforschungszentrum (MKFZ), Augustenburger Platz 1, 13353 Berlin, Germany

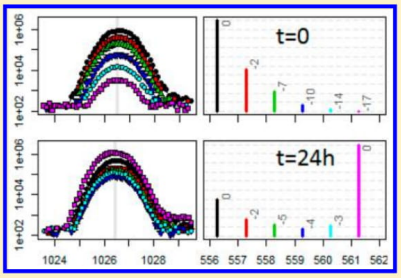
[‡]Berlin Institute of Health (BIH), Anna-Louisa-Karsch 2, 10178 Berlin, Germany

[§]Max-Delbrück-Center for Molecular Medicine (MDC), Robert-Rössle-Straße 10, 13125 Berlin, Germany

[∥]Federal Institute for Materials Research and Testing (BAM), Division 1.7 Analytical Chemistry, Richard-Willstätter-Straße 11, 12489 Berlin, Germany

Supporting Information

**ABSTRACT:** "Fluxomics" refers to the systematic analysis of metabolic fluxes in a biological system and may uncover novel dynamic properties of metabolism that remain undetected in conventional metabolomic approaches. In labeling experiments, tracer molecules are used to track changes in the isotopologue distribution of metabolites, which allows one to estimate fluxes in the metabolic network. Because unidentified compounds cannot be mapped on pathways, they are often neglected in labeling experiments. However, using recent developments in de novo annotation may allow to harvest the information present in these compounds if they can be identified. Here, we present a novel tool (HiResTEC) to detect tracer incorporation in high-resolution mass spectrometry data sets. The software automatically extracts a comprehensive, nonredundant list of all compounds showing more than 1% tracer incorporation in a nontargeted fashion. We explain and show in an example data set how mass precision and other filter heuristics, calculated on the raw data, can efficiently be used to reduce redundancy and noninformative signals by 95%. Ultimately, this allows to quickly investigate any labeling experiment for a complete set of labeled compounds (here 149) with acceptable false positive rates. We further re-evaluate a published data set from liquid chromatography-electrospray ionization (LC-ESI) to demonstrate broad applicability of our tool and emphasize importance of quality control (QC) tests. HiResTEC is provided as a package in the open source software framework R and is freely available on CRAN.




In analogy to other -omics technologies, the terms metabolomics and fluxomics are used to describe the investigation of metabolic levels and metabolic fluxes. Metabolite levels are usually measured in a static manner, e.g., at a given time point in a cell. While observable differences in metabolite levels are often highly informative, fluxes can be considered a more comprehensive way to describe cellular phenotypes as they represent a close functional link between all layers of cellular regulation.[1−3]

However, fluxes cannot be measured directly but must be calculated from the conversion rate of metabolites. To track the fate of metabolites and their dynamics in the metabolic network, stable isotope labeling experiments are conducted as the basis for fluxomics. The stable carbon isotope $^{13}C$, ubiquitously present in organic compounds, is most frequently used as a tracer molecule, although $^{15}N$, $^{18}O$, and $^{2}H$ are experimentally employed to a minor extent as well.[4]

To measure the abundance of small molecules, mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are most commonly used as analytical platforms. Coupling of gas chromatography (GC) and high resolution MS via soft ionization devices (GC-APCI-MS) not only increases sensitivity compared to traditional electron ionization (EI) but also facilitates elucidation of unknown compounds.[5−8] However, the use of GC requires analyte derivatization prior to measurement, which introduces significant amounts of nonbiological atoms into compounds, e.g., C, Si, O, and N in the case of methoximation/silylation. At the resolution of current high-resolution MS ($R \approx 50\,000$) isotopologues containing Si isotopes ($^{29}Si$ and $^{30}Si$) are not well separated from isotopologues containing $^{13}C$ on the mass scale but rather cause a measurable mass shift in the observed mass isotopomer distribution (MID). Data processing software needs to account for this mass shift and also for the presence of atoms of

**Received:** January 23, 2018
**Accepted:** May 25, 2018
**Published:** May 25, 2018





DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

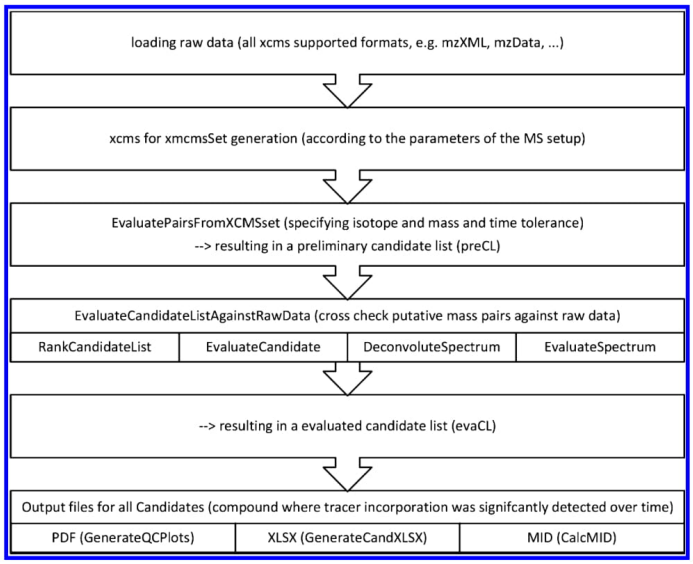



**Figure 1.** Flowchart of processing steps and functions in HiResTEC.

nonbiological origin. Tools developed so far for LC experiments generally do not meet these requirements, as derivatization is less common in LC.

Metabolomics and flux assays result in significant amounts of primary data rendering exclusive manual data curation barely feasible and calls for specialized automated data handling and evaluation tools. For evaluation of stable isotope labeling experiments a variety of software solutions are already published (geoRge,[9] MIA,[10] x13 cms,[11] NTFD,[12] iMS2Flux,[13] IsotopicLabelling,[14] mzMatch−ISO,[15] MetExtract,[16] and MetExtract II[17]), each addressing their own issues of partly very specific workflows and data quality or structure (ionization techniques (EI/ESI), LC/GC, targeted/nontargeted, nominal mass resolution, parallel labeling). However, none of these solutions is designed to investigate high-resolution GC-APCI data after derivatization in a nontargeted approach.

Here we present, describe and evaluate a novel tool for nontargeted, efficient, and reliable tracer incorporation detection in high-resolution mass spectra (HiResTEC), which is programmed in the open source language R and is freely available on CRAN. HiResTEC provides (semi-) automated quality check filters on sample raw data reducing the complexity of data sets by >95%, which allowed one to detect 149 unique labeled compounds in an example data set after visual inspection of the provided intuitive graphical output. An integrated novel experiment wide deconvolution algorithm is exploited to minimize redundancies and allow compound annotation. Testing for tracer incorporation builds internally on ANOVA linear models, which increases the statistical power to detect significant enrichments and allows one to analyze complex experimental setups in a highly flexible way. While developed for GC-APCI, HiResTEC also shows excellent performance on publicly available LC-ESI data.

## EXPERIMENTAL SECTION

**Sample Preparation for GC-APCI-MS Analysis.** All GC-MS samples evaluated with HiResTEC were processed as described in SI Text. In short, methanolic extracts were vacuum-dried and online derivatized using methoxyamine and *N*-methyl-*N*-(trimethylsilyl) trifluoroacetamide (MSFTA) before automatic injection by an RTC PAL system. In total >500 samples from 6 experiments investigating different cancer cell lines with various layouts (2−6 time points up to 24 h, [U]$^{13}$C-Glc and 1,2-$^{13}$C-Glc as tracer, 1−10 biological replicates) have been analyzed so far and allowed for heuristic evaluation of software performance. As one example set, we provide raw data for 36 samples originating from a lymphoma cell culture experiment, comprising two biological groups (G1 and G2) investigated at three time points ($t$ = 0 min, 10 min, and 24 h) and in 6 replicates per group and time point combination.

**Data Handling and Evaluation/Functions of the Package.** Measurement raw data can be exported in any file format supported by xcms. The algorithm is implemented as a set of R functions and can be installed as a package from CRAN. The principle workflow, and the function of the script are depicted in Figure 1. Besides the raw data files, information regarding sample time point (duration of labeling) and group (i.e., treatment, genotype or combinations thereof) is required. Example data were additionally preprocessed to correct for overloading peaks using the R package CorrectOverloadedPeaks.[18] The R package xcms is used to generate an xcmsSet object containing putative peaks within all samples (peak picking, grouping, and retention time alignment). Parameters were used as described elsewhere[9] (reanalysis of LC data) or using the following parameters for evaluation of GC-APCI example data set: method = "centWave", ppm = 25, peakwidth = c(1,6), snthresh = 1, prefilter = c(5,2000) and noise = 1 for


DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

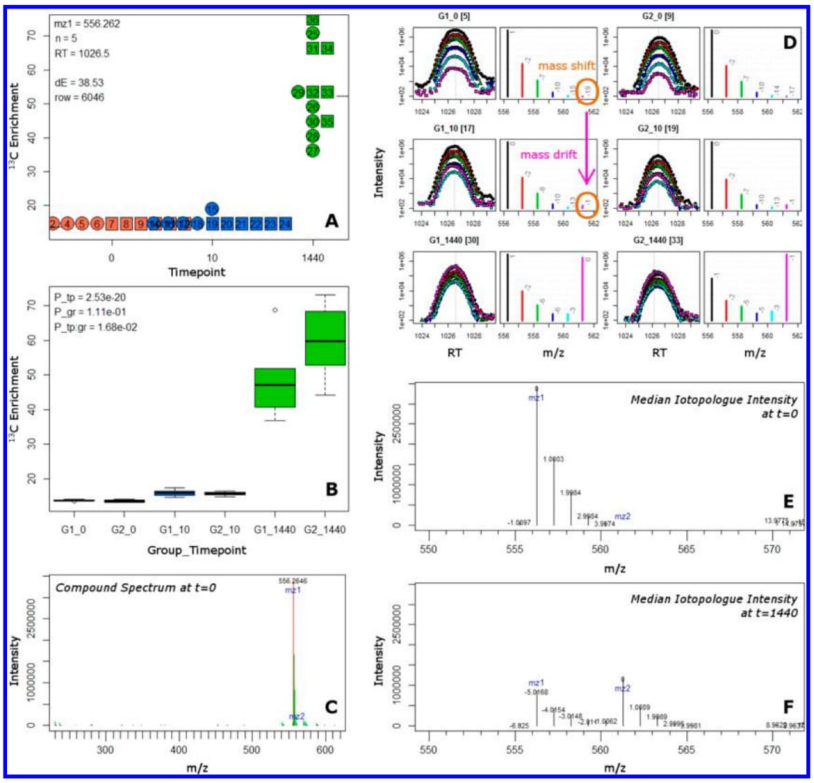



**Figure 2.** Graphical output (QC) for one compound (candidate 4, confirmed to be adenosine which contains a fully labeled ribose group). The subfigures focus on integrative data on enrichment (A, B), raw data information (D), and mass spectra (C, E, F). (A) Enrichment or relative $^{13}C$ amount per sample as calculated from MIDs depicted in part D. Plot symbols encode groups, e.g., treatment/control, symbol colors encoded time points and contain numbers to identify individual samples. Relevant information regarding masses and time are annotated at the top left together with a robust estimate of median enrichment change (d*E*) and the row number of preCL used as an input. (B) Boxplot version of the data from A amended by *p*-values of a linear model with factors time and group. (C) Spectrum as obtained by EWD for samples at $t_{initial}$. The current mass pair is highlighted (red) and annotated. Further masses present in the spectrum which are also found in preCL are highlighted in green. The spectrum can be used as a basis for compound annotation as well as to check if the best representative peak of a compound was selected. (D) Extracted BPCs for individual representatives of each group/time point-combination. The scan where the summed intensity over all investigated ions (from mz1 to mz1 + *n*) is at maximum, indicated by a gray line, represents the spectrum/MID plotted next to the BPCs and is used to calculate the enrichment for A. Numbers at each MID peak represent the mass shift, i.e., the difference of the measured mass and the theoretical $^{13}C$ isotope mass mz1+n*iso. As current time-of-flight-mass spectrometry (TOF-MS) with resolutions of about $R$ = 50 000 cannot resolve Si and C isotopologues, this mass shift is systematically negative for larger isotopologues at low labeling but will increase to zero in case of progressive $^{13}C$ incorporation (cf. SI Text). (E,F) Fraction of the spectra around the candidate peak pair deconvoluted from samples at $t_{initial}$ (E) and $t_{final}$ (F). Mass differences are labeled relative to mz1 and mz2, respectively. This subplot allows one to quickly access if the best mass pair was selected for this compound.

function xcmsSet and minsamp = 6, bw = 0.5 and mzwid = 0.25 for function group.

*EvaluatePairsFromXCMSSet.* In the next step, all peaks of the peak list will be combined pairwise depending on user specified thresholds for *n*, the number of maximum expected tracer atoms incorporated, *drt*, the maximum allowed retention time difference between two peaks in seconds (e.g., <2 s for well aligned GC-MS and higher for LC-MS data), and *dmz*, the maximum allowed mass deviation in Da (e.g., 0.004 Da for high precision instruments). The isotopic mass difference can be set as a parameter but has currently only been tested using 1.003355 Da for carbon labeling experiments. For all peak pairs, (relative) tracer incorporations and corresponding time dependent *p*-values are calculated based on group mean values and results are combined in a preliminary candidate list (preCL). The two masses defining a putative isotopologue-pair were arbitrary named as mz1 and mz2 for the smaller and larger detected isotopologue, respectively. Please note, this generally is not equivalent to M + 1 and M + 2 of the MID. We used the zz-index[19] with a cutoff of 0.3 to verify peak shape of sufficient quality in the base peak chromatograms (BPCs) of mz1 and mz2.



DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

**Table 1. Comparing Automatic Mass Pair Evaluation Using Raw Data BPCs against Simple xcms Peak Ratio Evaluation[a]**

| fraction | HiResTEC | xcms | description |
|---|---|---|---|
| *all* | 7462 | 2208 | number of mass pairs detected in preCL (HiResTEC) and number of significant results at $p < 0.05$ without raw data evaluation (xcms) respectively |
| **tested* | 5947 | 1432 | evaluated mass pairs from preCL: evaCL |
| ***candidates* | 347 | 169 | positively evaluated mass pairs |
| ***rejected* | 5600 | 1263 | rejected mass pairs violating QC rule |
| **untested* | 1515 | 776 | untested mass pairs (redundant information) |

[a]Of *all* observed mass pairs ($n$ = 7562), most will be subjected to automatic evaluation against raw data (*tested*) while others remain *untested* as they correlate with a positive candidate. Positively tested pairs become *candidates* and negatively tested get *rejected*. If no raw data evaluation is conducted, candidates could be obtained by evaluating xcms reported intensities directly. Out of all 7462 peak pairs, this would lead to 2208 candidates at $P < 0.05$. However, these contain redundancy ($n$ = 776) and false positives ($n$ = 1263). Column xcms indicates the number of peak pairs from this approach without QC.

(correlated to an already positively evaluated pair, $n$ = 1515) or because they did not fulfill all of the QC ($n$ = 5600).

The importance of the second step (evaluation against raw data) becomes apparent when results are compared against such obtained from peak intensity data evaluation alone (first step). Out of the 2208 significant pairs, only 169 are found in the candidate list after step two. The remaining 92% represent either redundant information or were rejected during QC testing. One reason for the large fraction of rejected pairs which appeared significant in first analysis is that sensitive peak detection settings, necessary to identify low abundant isotopologue intensities, can result in peak artifacts in noisy chromatographic regions.[21] QC plots of all accepted and selected rejected candidates are provided in SI File 1 and SI File 2, respectively.

Automatic QC criteria do not exclude False Positive (FP) candidates completely. Manual inspection, using automatic produced QC plots, described in detail further below, allowed to reject further 198 of the 347 positively evaluated candidates, resulting in 149 TPs equivalent to an overall FPR of 43% (SI Table 1). While this FPR may seem high, it has to be put into perspective of the much higher number of automatically rejected peaks ($n$ = 5600) whose manual curation would have been impossible. Further, FP occurrence is linked to peak intensity, with smaller peaks more prone to be detected as FP. As peak pairs are ranked according to intensity before QC testing and hence large peaks are evaluated first (cf. Experimental Section), this leads to smaller FPR among more abundant peaks when considered separately (FPR = 0% in top 50 candidates, 11% in top 100, 26% in top 200..., SI Table 1). Thus, it remains the decision of the user to balance off work load of manual curation against potentially overlooking minor enriched peaks.

**Mass Drift Is an Important Filtering Heuristic Allowing for Excluding False Positives.** Several quality criteria were implemented as automatic tests to evaluate each candidate. Results of these tests are summarized in SI Text and examples regarding each error message can be found in SI File 2. As expected, many candidate peak pairs were rejected because of insignificant $P$-values ($P$ > 0.01: 73%) and/or a negligible change in enrichment (d$E$ < 1: 62%). Mass shifts/drifts outside of the expected range were observed in 40% of all candidates, rendering it an important QC in the present nontargeted approach. To further test this, we evaluated the test data set without the mass drift filter being activated (column "Test" in Table 1). This resulted in >120 additional candidates which were almost exclusively False Positives.

Changes in mass between samples occur because sample derivatization in GC-MS leads to defined chemical modifications of analytes and ultimately adds carbon and silicon atoms of nonbiological origin to the molecules. In short, two effects take place: (i) isotopologues of a compound in nonlabeled samples are found at masses lower than expected from the calculation of mz1+n*1.00335, i.e., showing a detectable mass *shift* and (ii) isotopologues of a compound in labeled samples approach the expected value, i.e., do not show a mass shift. As a consequence, for the same measured isotopologue, mass shift values will differ between labeled and nonlabeled states. As differences in labeling occur over time, we termed this effect mass *drift*. We observed many false positive candidates due to coeluting peaks being produced or consumed during the experiment which can be detected by mass drift QC. Additionally, we observed in QC plots that this effect is sensitive enough to allow to detect tracer incorporation before significant changes in intensity are apparent (cf. SI File 1, page 90 for an example where the M + 3 isotope is only minimal increased in intensity but shows a strong mass drift of 5 mDa comparing $t$ = 0 and $t$ = 24 h). A more detailed explanation off mass shift and mass drift can be found in SI Text.

**Use of ANOVA Models and Experiment Wide Deconvolution Facilitates Nontargeted Analysis.** Usually flux experiments run nonlabeled control samples in parallel, comparing changes in sample subsets on, e.g., a time point by time point basis. In HiResTEC, we implemented an ANOVA based approach which allows one to track changes in enrichment using two or a variable higher number of time points. If annotation of compounds is of lesser importance, nonlabeled samples (e.g., control at $t$ = 0) can be omitted altogether. This approach allows one to incorporate any grouping factor (e.g., genotype, treatment, or a combination thereof) to be included in the analysis, thereby increasing the statistical power, allowing for flexible experimental setups and avoiding time extensive subgroup analysis.

We utilize mass spectra extracted from raw data by EWD for three distinct purposes: redundancy removal, quality control and annotation. Most compounds will be represented in preCL multiple times due to in-source fragments and adducts as well as various $m/z$-pair combinations within the same MID (e.g., we would find 3 independent pairs in cases of 3 measured isotopologues: M + 0/M + 1, M + 0/M + 2, and M + 1/M + 2). If a mass pair is processed and passed all QC, all further pairs detected at this RT and including masses which are present in the candidate's spectrum can be removed without further testing. This consequently saves processing time and limits the final output to the relevant information. A prerequisite for this approach is to test the best representative candidate pair from a spectrum first (as it would be removed and not tested if a less optimal pair is tested prior). This is achieved by a sorting heuristic described in the Experimental Section (*RankCandidateList*), which ensures that the base peak of a spectrum together with the M + $n$ showing the strongest intensity increase over time are ranked highest.



DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

*EvaluateCandidateListAgainstRawData.* The candidate pairs from preCL are tested against raw data. About 2000 candidate pairs can be processed in 15 min on a modern desktop computer. The aim is to perform a number of tests on raw data chromatogram traces as well as on putative spectra to ultimately assign informative error messages and reject the candidate or to include it in a nonredundant evaluated candidate list (evaCL). Several steps proved valuable to achieve this and are described in the following. An overview of all criteria and default cutoff values are listed in SI Text.

*RankCandidateList.* Every compound can be expected to appear several times within preCL due to multiple isotopologues, fragments, and adducts, which are all present in the compound's (pseudo) mass spectrum. Ideally, we only need to test the best representative $m/z$-pair out of such a spectrum, where best means to preferably include a high intensity peak (e.g., base peak) and showing strong tracer incorporation. *RankCandidateList* sorts preCL according to summed intensity of the mz1/mz2 pairs over time, which allows to test the most promising candidates first and speed up processing by stepwise removal of all redundant pairs from preCL belonging to the same pseudospectrum (cf. *DeconvoluteSpectrum*).

*EvaluateCandidate.* For all relevant ions at a specific retention time (RT) and framed by a candidate peak pair mz1 and mz2, where mz2 = mz1+n*isotopic mass, BPCs are extracted from all samples depending on user specified parameters (drt, dmz, method). The relative $^{13}C$ tracer incorporation (enrichment $E$) is calculated by the amount of $^{13}C$ divided by total C in the mass peak[20] over all isotopologues from mz1 to mz2.

$E$ is further evaluated in a linear model incorporating time (TP) and group (GR) information as well as the interaction thereof. We expect TP as numeric with at least 2 unique values (initial and final), but several TP and even the omission of TP = 0 can be analyzed. We allow an additional grouping factor to be included which may contain genotype, treatment, or other information, possibly combined if more than 2 levels are given. Time related ANOVA $P$-values from such a model together with the change of enrichment over time (d$E$) are calculated and used together with median peak intensity as QC filter criteria applying user defined cut-offs.

*DeconvoluteSpectrum and EvaluateSpectrum.* An experiment wide deconvolution (EWD) function is internally utilized to extract the mass spectrum containing a specified target mass (mz1) at a certain RT. To this end, for appropriate statistics a minimum of 5 raw data files at $t_{initial}$ need to be provided together with retention time information on a specific peak as well as allowed deviation values for mz and RT. EWD will then extract the maximum intensity ($I$) of BPCs for all candidate masses coeluted with mz1 and their precise RTs for all provided raw data files within the allowed deviations. In the next step, each of these candidate masses will be tested for colocalization of the apexes and Pearson correlation of intensities over all data files. For GC-APCI data recorded at a scan rate of 10 Hz we found dRT = 0.15 s and $r$ = 0.7 to be suitable cut-offs to decide which candidate masses are to be included in the final mass spectrum. To finally reconstitute an average spectrum, intensity ratios are determined in a robust manner by calculating the median over all samples and scaled according to the intensity of the data file where $I_{mz1}$ is at maximum.

Following EWD the data quality in the spectrum is rated by several measures. Requirements are, for example, that the candidate under evaluation is not below 10% intensity of the base peak in the spectrum to avoid spurious signals from minor fragments. Further, suggested candidates have to include the peak of highest intensity in the MID which helps to exclude false positives due to $[M^+]$ ions or isotopologue pairs not including the [M + H] and the strongest labeled fragment.

*Various Output Functions. EvaluateCandidateListAgainstRawData* produces a list of evaluation results (evaCL) containing both, all nonredundant candidate pairs, and all rejected candidates. The information from evaCL can be written into graphical output, i.e., QC-plots as a pdf file for the visual examination (Figure 2), an xlsx spreadsheet listing all compounds, or can be used to generate natural abundance corrected MIDs.

*Systematic Performance Evaluation against a Public Data Set.* To test the performance of HiResTEC on LC data, the script was evaluated against and compared to the data set provided as Supporting Information in Capellades et al.[9] We followed two approaches to generate the preCL. First, we converted "IncIDs" reported by Capellades into our preCL format. Second, we reprocessed the provided mzData files using the reported parameters to generate an xcmsSet and, in the following, used our own function to generate a preCL from this xcmsSet. Both preCL versions were then evaluated against raw data before comparing the results.

## RESULTS AND DISCUSSION

The aim of this work was to provide a software solution to find all compounds which had incorporated tracer molecules in a nonredundant fashion to assist in data preparation for flux modeling tools. None of the available tools (for a tabular comparison see Table S3 in SI Text) provided satisfying results. This is in part because appropriate treatment of mass shifts (explained further below) was not implemented. We therefore expanded the two stage strategy implemented, e.g., by Capellades et al.[9] in the software geoRge, of first detecting all peaks in each sample and second evaluating potential candidate pairs by mass difference analysis statistically. It soon became obvious that this approach did yield vast numbers of false positives results, i.e., due to the presence of noise peaks, which were found by manual inspection. We, therefore, implemented a series of heuristic tests to identify false positives automatically and then developed an intuitive plot layout for manual quality control both of which are explained in the following. For demonstration purposes we analyzed a data set of 36 samples evaluating 2 types of cells at 3 time points using 6 replicates per cell type/time point combination.

**Example Data Set Evaluation Reveals 149 Nonredundant Compounds Showing Significant Tracer Incorporation and Rejecting Automatically 95% of Initial Candidates.** Following the two step procedure described in detail in the methods section, we evaluated all 36 samples combined for the presence of ion peaks and combined these ions into pairs if they were colocated and their mass difference was a multiple of 1.00335 Da, the mass difference of $^{13}C$ and $^{12}C$. In total, 7462 of such pairs existed after step one (Table 1, equivalent to preCL in Figure 1), 2208 of which showed significant changes in peak intensity ratio over time, indicative of cumulative tracer incorporation. In step two, we evaluated all 7462 pairs against raw data. This resulted in 347 nonredundant positively evaluated mass pairs, each defining a MID and respective enrichment which significantly changed over time. In total, automatic evaluation removed >95% of all pairs either because they were redundant



DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

EWD spectra are an essential part of QC plots and can be used to intuitively check if the candidate really represents the base peak of the spectrum (Figure 2C) and if the selected M + *n* is the optimal choice (Figure 2E). As spectra will change over time (due to tracer incorporation), EWD is performed twice on different subsets of raw data files which either contain only samples of the initial time point (e.g., $t = 0$, all spectra similar between different groups) or samples from the final time point (spectra may show inter group variance). Further, spectra can be evaluated for the mass pair intensity relative to the base peak and if it includes the main peak of the MID, which are both criteria used to reject potential candidates (SI Table 2). The EWD spectra additionally serve as a basis for putative identification. We have previously demonstrated that interpretation of in-source fragments tremendously improves sum formula prediction for GC-APCI spectra.[22]

**Visual QC Output Allows for Fast Manual Curation of Candidates and Confirms Manageable Amount of Remaining False Positives.** We implemented a rigorous evaluation of raw data BPCs for all relevant ions to verify true observations and exclude redundant or incorrect results. Ideally, a true positive compound should present significant and systematic tracer incorporation/enrichment increase (Figure 2A,B), chromatographically correlated isotopologue peaks with predictable mass shifts (Figure 2D), and an interpretable spectrum of fragments and adducts (Figure 2C).

These plots can be automatically generated for all or subsets of evaluated peak pairs (evalCL) and serve as a basis to estimate the false positive rate (FPR). We allow for relatively high FPRs by choosing permissive cut-offs to avoid a larger number of false negative results since high FPR in ∼350 candidates is possible to manually control by visual data inspection while scanning >5000 rejected candidates for FNs is prohibitive. Example files presenting all automatically evaluated positive and selected negative candidates are provided as SI File 1 and SI File 2 to give a broader overview of observed results. Selection in SI File 2 was done randomly and limited to 100 candidates to keep the file size small. The complete PDF (>100 MB) can be generated from the example data or obtained from the authors.

Manual data inspection is preferred over automatic filters in evaluating properties of the extracted ion chromatograms (like bad peak shape or coeluting compounds), which are often easier identifiable by eye than algorithmically. To a minor extent, unusual spectra properties or strong variance in obtained enrichment values are a good guide to decide if the compound at hand is a FP or not. In SI Table 1 rejected peaks were roughly classified with a letter code indicating the reason for rejection. The majority (101/149) are classified as questionable or of neglectable effect. A good example here would be candidate 325 (p.1297ff). BPCs indicate mz1 and mz2 to be valid peaks with the average increase in enrichment over time (d*E*) being positive (7.8%). However, the variance within groups (∼10%) is larger than the detected effect, rendering it of less importance or potentially a FP. Besides, the deconvoluted spectra is very sparse, which will hamper putative annotation. Another example, fulfilling all automatic QC tests but rejected for similar reasons would be candidate 344 (p.1373ff). Here, additionally not even a clear peak can be detected by eye in the BPC plot. However, also large peaks with complex spectra were rejected. Candidate 102 (p.405ff), for example, shows only minor tracer enrichment (2.85%) compared to the variance within sample groups. Further, in the mass spectrum, the green highlighted peaks, indicate that many other peak pairs of higher intensity were already subject to automatic evaluation and obviously rejected, because otherwise this peak would have been removed from the candidate list (preCL) due to spectral correlation. Candidate 190 (p.758ff) would be an example for suspicious results due to a coeluting peak easily observable in the BPC plot.

**Systematic Evaluation of HiResTEC on a Published LC Data Set.** Capellades et al. published a conceptually similar software solution geoRge, analyzing LC-MS data nontargeted for tracer incorporation by investigating intensity ratio changes in xcms derived peak lists.[9] Applying geoRge on APCI data we observed a huge number of false positive results when compared against raw data. Causal were xcms peak detection artifacts, strong redundancy, and large mass shifts in APCI peaks between time points. In consequence, this unfortunately prevented to successfully process APCI data sets with geoRge. To test if HiResTEC is also capable to analyze LC-MS data we re-evaluated the example data provided by Capellades systematically in an attempt to reproduce published results.

First, we used the provided raw data and xcms parameters to calculate the xcms peak lists and the potential mass pairs de novo (20099 peak pairs). Second, we reconstructed a mass pair list from the reported 271 "IncIDs" found in Supporting Information material with all corresponding isotopologues (690 peak pairs). Both lists were subjected to QC testing against raw data in the following and results compared regarding their overlap (Figure 3).

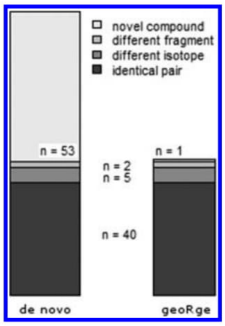



**Figure 3.** Remaining compounds with significant tracer incorporation in both data sets. Column de novo accounting for all hits from the forward approach, reprocessing the raw data files and geoRge accounting for all hits in the backward approach, re-evaluating the reported hit list, and the corresponding sub classification.

From the reconstructed list, 48 nonredundant compounds showing significant tracer incorporation were obtained. This number is much lower than the Inc.IDs reported due to removed redundancy and automatic rejection of some candidates following raw data evaluation (cf. SI File 3 and SI File 4 for accepted and rejected candidates, respectively). All but one of these 48 compounds are also found when processing a preCL de novo. However, in five cases a different isotope was selected and in two cases a different fragment from the mass spectrum of the compound (Figure 3). Strikingly, de novo evaluation allowed one to detect 53 additional, previously unreported compounds (SI File 5). Manual inspection of QC plots revealed a FPR of 12.5% in the 47 overlapping and 39.5%


DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

in the novel compounds, respectively (SI Table 2). We also determined the False Negative Rate in the rejected mass pairs (SI File 4) to be <3%.

In LC-MS the above-described mass drift is not present, but nevertheless mass shift, if present, often indicates coeluting compounds and can be used for QC purposes. In total, 11% of all rejected candidates show this error message. The intensity threshold (parameter Icut = 2500) was found to be a useful filter criterion to identify False Positives in LC-MS data. While being of minor importance in the GC-APCI test data set it allowed one to reject 58% of all peak pairs in LC-ESI. Peaks not exceeding at least 2500 intensity counts in at least half of all samples for mz1 at $t = 0$ or mz2 at $t = 24$ were often found to not show satisfying peak shapes and give rise to spurious results. In conclusion, reprocessing a publicly available LC-ESI data set using additional raw data QC analyses allowed one to remove a large amount of redundancy, reject several FPs, and detect additional compounds showing tracer incorporation.

**Analysis Results Can Be Used for Multiple Purposes.** The main function of HiResTEC returns a full list of all evaluated peak pairs which may serve as a basis for further analyses like evaluation of filter performance, annotation of spectra, preparation of a reference list or library file and export of MIDs for flux modeling software. Export functions are provided for tabular output (Excel spreadsheet) and various figure formats (PDF, similar to Figure 2). The corresponding mass spectra are saved together with all extracted BPCs and derived statistical data and can be subjected to tools like *InterpretMSSpectrum*[23] for sum formula assignment and annotation. Following compound identification, MIDs can be corrected for natural isotopic abundancies in all relevant compounds as a preparation for flux modeling software tools.

## CONCLUSIONS

We developed a practical software solution for the sensitive and nontargeted detection of tracer enrichment in respective high-resolution GC- and LC-MS data sets. Applying a two-step procedure of peak picking and evaluation against raw data BPCs, we reduce, through a set of automatic QC filters, redundancy and irrelevant data by about 95%. The results can be efficiently inspected, exported, or further analyzed by provided functions. The tool is freely available as an R package (HiResTEC) from CRAN (https://cran.r-project.org/) and can be incorporated in a fluxomics data processing pipeline.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b00356.

SI Table 1, evaluated candidate list of example data including annotation of likely False Positives due to inspection of QC plots from SI File 1 (XLSX)
SI Table 2, evaluated candidate list of reprocessed LC-MS data including annotation of likely false positives due to inspection of QC plots from SI File 3 (XLSX)
SI File 1, all QC plots for accepted candidates of ExampleData (PDF)
SI File 2, 100 randomly selected QC plots for rejected candidates of ExampleData (PDF)
SI File 3, all QC plots for accepted candidates of reanalyzed LC-MS data (geoRge) (PDF)
SI File 4, all QC plots for rejected candidates of reanalyzed LC-MS data (geoRge) (PDF)
SI File 5, all QC plots for accepted candidates of reanalyzed LC-MS data (de novo) (PDF)
SI Text, detailed information on QC filters and function parameters, GC-APCI processing and the mass drift phenomenon, and comparison of currently available tools for tracer incorporation detection (PDF)

## AUTHOR INFORMATION


### Corresponding Author

*E-mail: jan.lisec@bam.de. Fax: +49 (30) 8104-75891.

### ORCID

Jan Lisec: 0000-0003-1220-2286


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

J.L. received funding from Berlin Institute of Health, and F.H. received funding from Berlin School of Integrated Oncology.

## REFERENCES


(1) Winter, G.; Krömer, J. O. *Environ. Microbiol.* **2013**, *15* (7), 1901–1916.
(2) Duckwall, C. S.; Murphy, T. A.; Young, J. D. *J. Carcinog.* **2013**, *12*, 13.
(3) Moreno-Sánchez, R.; Saavedra, E.; Gallardo-Pérez, J. C.; Rumjanek, F. D.; Rodríguez-Enríquez, S. *FEBS J.* **2016**, *283* (1), 54–73.
(4) Klein, S.; Heinzle, E. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2012**, *4* (3), 261–272.
(5) Wachsmuth, C. J.; Hahn, T. A.; Oefner, P. J.; Dettmer, K. *Anal. Bioanal. Chem.* **2015**, *407* (22), 6669–6680.
(6) Wachsmuth, C. J.; Almstetter, M. F.; Waldhier, M. C.; Gruber, M. A.; Nürnberger, N.; Oefner, P. J.; Dettmer, K. *Anal. Chem.* **2011**, *83* (19), 7514–7522.
(7) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, *9*, 44–66.
(8) Strehmel, N.; Kopka, J.; Scheel, D.; Böttcher, C. *Metabolomics* **2014**, *10* (2), 324–336.
(9) Capellades, J.; Navarro, M.; Samino, S.; Garcia-Ramirez, M.; Hernandez, C.; Simo, R.; Vinaixa, M.; Yanes, O. *Anal. Chem.* **2016**, *88* (1), 621–628.
(10) Weindl, D.; Wegner, A.; Hiller, K. *Bioinformatics* **2016**, *32* (18), 2875–2876.
(11) Huang, X.; Chen, Y.-J.; Cho, K.; Nikolskiy, I.; Crawford, P. A.; Patti, G. *J. Anal. Chem.* **2014**, *86*, 1632–1639.
(12) Hiller, K.; Metallo, C. M.; Kelleher, J. K.; Stephanopoulos, G. *Anal. Chem.* **2010**, *82* (15), 6621–6628.
(13) Poskar, C. H.; Huege, J.; Krach, C.; Franke, M.; Shachar-Hill, Y.; Junker, B. H. *BMC Bioinf.* **2012**, *13* (1), 295.
(14) Ferrazza, R.; Griffin, J. L.; Guella, G.; Franceschi, P. *Bioinformatics* **2017**, *33* (2), 300–302.
(15) Chokkathukalam, A.; Jankevics, A.; Creek, D. J.; Achcar, F.; Barrett, M. P.; Breitling, R. *Bioinformatics* **2013**, *29* (2), 281–283.
(16) Bueschl, C.; Kluger, B.; Berthiller, F.; Lirk, G.; Winkler, S.; Krska, R.; Schuhmacher, R. *Bioinformatics* **2012**, *28* (5), 736–738.
(17) Bueschl, C.; Kluger, B.; Neumann, N. K. N.; Doppler, M.; Maschietto, V.; Thallinger, G. G.; Meng-Reiterer, J.; Krska, R.; Schuhmacher, R. *Anal. Chem.* **2017**, *89*, 9518–9526.
(18) Lisec, J.; Hoffmann, F.; Schmitt, C.; Jaeger, C. *Anal. Chem.* **2016**, *88* (15), 7487–7492.
(19) Zhang, W.; Zhao, P. X. *BMC Bioinf.* **2014**, *15*, S5.
(20) Fisher, T. R.; Haines, E. B.; Volk, R. *J. Limnol. Oceanogr.* **1979**, *24* (3), 593–595.




DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253–7260

(21) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. *Anal. Chem.* **2017**, *89*, 8689−8695.
(22) Jaeger, C.; Hoffmann, F.; Schmitt, C. A.; Lisec, J. *Anal. Chem.* **2016**, *88*, 9386−9390.
(23) Jaeger, C.; Méret, M.; Schmitt, C. A.; Lisec, J. *Rapid Commun. Mass Spectrom.* **2017**, *31* (15), 1261−1266.



DOI: 10.1021/acs.analchem.8b00356
*Anal. Chem.* 2018, 90, 7253−7260

## Curriculum Vitae

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

# Publikationsliste

*Optogenetic monitoring identifies phosphatidylthreonine-regulated calcium homeostasis in Toxoplasma gondii.*

Kuchipudi A, Arroyo-Olarte RD, Hoffmann F, Brinkmann V, Gupta N.
Microb Cell. 2016 May 2;3(5):215-223. doi: 10.15698/mic2016.05.500.
PMID: 28357357

*Extending the Dynamic Range in Metabolomics Experiments by Automatic Correction of Peaks Exceeding the Detection Limit.*

Lisec J, Hoffmann F, Schmitt C, Jaeger C.
Anal Chem. 2016 Aug 2;88(15):7487-92. doi: 10.1021/acs.analchem.6b02515. Epub 2016 Jul 21.
PMID: 27377477

*Automated Annotation and Evaluation of In-Source Mass Spectra in GC/Atmospheric Pressure Chemical Ionization-MS-Based Metabolomics.*

Jaeger C, Hoffmann F, Schmitt CA, Lisec J.
Anal Chem. 2016 Oct 4;88(19):9386-9390. Epub 2016 Sep 14.
PMID: 27584561

*Nontargeted Identification of Tracer Incorporation in High-Resolution Mass Spectrometry.*

Hoffmann F, Jaeger C, Bhattacharya A, Schmitt CA, Lisec J.
Anal Chem. 2018 Jun 19;90(12):7253-7260. doi: 10.1021/acs.analchem.8b00356. Epub 2018 Jun 7.
PMID: 29799187

*Pharmacokinetics of Daclatasvir, Sofosbuvir and GS-331007 in a Prospective Cohort of HCV positive Kidney Transplant Recipients.*

Schrezenmeier E, Hoffmann F, Jaeger C, Schrezenmeier J, Lisec J, Glander P, Algharably E, Kreutz R, Budde K, Duerr M, Halleck F.
Ther Drug Monit. 2018 Oct 30. doi: 10.1097/FTD.0000000000000567. [Epub ahead of print]
PMID: 30422962

# Danksagung

Mein besonderer Dank gilt zunächst Dr. Jan Lisec und Dr. Carsten Jaeger für das Heranführen an Massenspektrometrie und deren Analysenmethoden, sowie unzählige Programmiertutorials und debugging-Hilfen. Im Besonderen bedanke ich mich auch für euer mir entgegengebrachtes Vertrauen und Geduld, im Labor, beim Analysieren, Schreiben und Testen.

Für konstruktive Anregungen danke ich ebenso bei meinem Gutachter-Komitee, bestehend aus Prof. Dr. Clemens Schmitt, Prof. Dr. Klaus Irrgang und Prof. Dr. Michael Schupp, die meine Doktorarbeit als weitere Gutachter betreut haben.
Dank gilt auch der Berlin School of Integrative Oncology ohne deren finanzielle, wissenschaftliche und organisatorische Unterstützung, die vorliegende Studie nicht möglich gewesen wäre.

Danken möchte ich außerdem meinen Mitstudenten der BSIO, im Besonderen Friederike Christen, Kaja Hoyer und Raphael Hablesreiter, die mich begleitet und moralisch unterstützt haben.

Schließlich, gilt mein besonderer Dank neben vielen Freunden meinen Eltern, Kathrin und Armin Hoffmann, die meine Arbeit mit großem Engagement unterstützt haben. Der größte Dank gebührt Baruch Francisco Velez-Rodriguez, der mich bis zur Vollendung dieser Arbeit unermüdlich unterstützt und ermutigt hat.