Technische Universität München

Automated annotation of structurally uncharacterized metabolites in human metabolomics studies by systems biology models

Jan-Dominik Bernd Quell

2019

Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Technische Universität München

Lehrstuhl für Experimentelle Bioinformatik

Automated annotation of structurally uncharacterized metabolites in human metabolomics studies by systems biology models

Jan-Dominik Bernd Quell

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Dmitrij Frischmann
Prüfer der Dissertation:	1. Prof. Dr. Hans-Werner Mewes
	2. Prof. Dr. Bernhard Küster

Die Dissertation wurde am 28.11.2018 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 28.01.2019 angenommen.

„Der Mensch muß bei dem Glauben verharren, daß das Unbegreifliche begreiflich sei; er würde sonst nicht forschen.“

Johann Wolfgang von Goethe

Danksagung

Zunächst möchte ich mich ganz besonders bei meinem Doktorvater (Hans-)Werner Mewes bedanken, der mir nicht nur die Möglichkeit gegeben hat an seinem Institut des Helmholtz Zentrums München sowie seinem Lehrstuhl an der Technischen Universität München in Forschung und Lehre zu arbeiten, sondern auch meine Laufbahn ab dem ersten Tag meines Studiums sowohl im fachlichen als auch im persönlichen Kontext entscheidend geprägt hat.

Mein besonderer Dank gilt auch Gabi Kastenmüller. Ihr möchte ich für viele fachliche Ideen, Diskussionen, und Inspirationen danken, durch die wichtige Weichen gestellt wurden und in meinen Projekten mündeten.

Für fachliche Diskussionen insb. mit biochemischen Inhalten und für die Unterstützung bei massenspektrometrischen Messungen zur Verifikation von Kandidatenmolekülen im Labor bedanke ich mich sehr herzlich bei Werner Römisch-Margl.

Da ich in dieser Arbeit Daten aus zahlreichen Kohortenstudien verwendet habe (u.a. SUMMIT, KORA, HuMet) geht an dieser Stelle ein Dank an meine Kooperationspartner wie Anna Köttgen (GCKD) und Karsten Suhre (QMDiab).

Bedanken möchte ich mich auch bei Bernhard Küster für die fachliche Auseinandersetzung mit meiner Arbeit und die Bereitschaft als zweiter Prüfer zu fungieren.

Natürlich geht auch ein Dank an meine Kollegen, Freunde und Familie, die mich auf unterschiedliche Arten und Weisen unterstützt haben. Die Vollendung dieser Arbeit markiert den Abschluss eines Kapitels in meinem Leben; nun freue ich mich auf neue berufliche Herausforderungen.

Nicht nur aus Tradition gebührt zum Abschluss ein sehr großer Dank meiner Frau Christina. Ich möchte ihr Danke sagen für viele Formen der Unterstützung bei meiner Arbeit, ebenso bei dem Verwirklichen von privaten Ideen und Zielen und für gemeinsam geschmiedete Lebenspläne, die erst durch sie Bereicherung finden.

Abstract

Metabolomics studies are commonly performed in research of health and disease, capturing metabolite levels as intermediate phenotypes that bridge interacting biochemical levels from the genome to clinically relevant phenotypes and reflect environmental influences. While defined sets of metabolites are quantified by targeted mass spectrometry (MS) for testing of hypotheses, non-targeted MS includes measurement of metabolites with initially unknown chemical structures and enables hypothesis-free testing of unexpected associations. Although metabolites without any chemical characterization only occur in non-targeted settings, targeted measurements can lead to ambiguous annotations too. The outcome of metabolomics studies strongly depends on the insights that can be obtained from measured data: Functional or intermediate metabolite features influence measurements of highly sensitive mass spectrometers, which differ in their mass resolution (e.g. ± 0.3 or ± 0.01) and in their metabolite separation capacity. Although separation of molecules can be improved by a coupled chromatography or multiple reaction monitoring (MRM), some molecules that share similar chemical or physical properties are therefore not separated analytically and labeled ambiguously (measurement of metabolite sums, e.g. hexoses, steroids, or complex lipids). As unknown or ambiguously labeled metabolites in scientific results mark dead ends in metabolomics research, they cannot be interpreted in concrete biochemical contexts, or further used in clinical applications. In consequence, unknown or ambiguously labeled metabolites need to be characterized precisely. Experimental approaches in wet-laboratories are successful in identification of individual unknown metabolites, but they are comparably time-consuming and laboratory-intensive. For metabolite annotation in large scales, high-throughput *in silico* approaches that mostly analyze fragmentation spectra have been introduced in recent years. Since some commercial MS platforms do not provide spectra and the development of MS devices leads to faster measurements, larger data sets, and increasing numbers of metabolites (also of those with unknown structural identities), characterization of unknown metabolites will stay a long-term accompanying future challenge.

Here, I propose two novel automatic approaches for (*i*) determination of main constituents of measured metabolite sums and for (*ii*) characterization of unknown metabolites including automated selection of candidate molecules relying mostly on non-spectral data. In the first part, I provide an automatic workflow to determine main constituents of ambiguously annotated metabolites based on a platform comparison in a small data (sub-) set and impute concentrations of a (large) number of samples on a higher precision level: In case of phosphatidylcholines (PCs) that consist of a polar head group and two non-polar fatty acid chains, most established lipidomics platforms measure the total length of side chains including their desaturation but cannot distinguish the distribution to the *sn*-1 or *sn*-2 positions (lipid species level). Recent platforms identify the length and desaturation of both side chains separately, but do not provide their ordering, or position of double bonds (higher precision; fatty acid level). To this end I used 76 and 109 PC concentrations measured by Absolute*IDQ*$^{\text{TM}}$ that has frequently been used during numerous large cohort studies for targeted quantification of amino acids and lipids on the lipid species level, and by the more precise Lipidyzer$^{\text{TM}}$ platform (fatty acid level), respectively, in the same 223 samples of healthy young men. First, I determined 150 to 456 theoretically possible constituents (fatty acid level) for 76 measured PC sums (lipid species level) by systematically distributing their fatty acid chains and desaturation to the *sn*-1 and *sn*-2 positions. 69 of 109 PCs were identified as constituents for 38 PC sums, of which 10 compositions contained one clear main constituent with a contribution of at least 80%. Further 17 compositions consisted of one or two measured main constituents with contributions between 20 and 80%. In one case, a PC with one odd-chain fatty acid was identified as main constituent, which was an unexpected finding, because in human matrices even-chain fatty acids are much more abundant than odd chain lengths. Since 13 and 19 quantitative compositions, respectively, were replicated directly, or via imputation of metabolite concentrations in samples of a second independent cohort, determined factors can be used for imputation of PC concentrations (lipid species level → fatty acid level). Although PCs measured on the fatty acid level are not equal to chemical structures, they allow more precise interpretations in their biochemical context.

In the second part, I introduce a systems biology model that approaches characterization of unknown metabolites, measured by non-targeted MS, from a holistic perspective. While current methods are focused on transforming and linking fragment spectra of unknown structures to database compounds, my approach identifies the biochemical context of unknown metabolites, before selecting specific candidate molecules. On this account, I use 637 concentrations (388 known, 249 unknown) measured in 2 279 samples to estimate pairwise partial correlations that are known to reconstruct biochemical pathways. The automatic data integration module merges significant associations with a published Gaussian graphical model (GGM) to receive 1 040 edges as backbone of my network model. By mapping and adding metabolites, I integrated 299 metabolite-gene associations of a published mGWAS, and 1 152 reactions as well as 495 functional metabolite-gene relations of the public database Recon. Several metabolite characterization modules that access the network model and transfer information from known to unknown metabolites predicted pathways for 180 unknown metabolites and selected 21 reaction types for 100 pairs of known and unknown metabolites. 12 reactions were substantiated by a second criterion on the retention time. I replicated 18 of 26 pathway annotations that were independently predicted in two data sets, based on high-resolution metabolite levels measured in 1 209 samples of a second independent cohort. For 139 of 315 connected unknown high-resolution metabolites that are connected to known metabolites in the network model, I automatically selected 676 candidate molecules by searching public databases for structurally similar compounds (to neighboring known metabolites) with equal mass. Finally, I verified two out of five selected commercially available candidates experimentally in a wet-laboratory and propose a database structure to organize several predicted or verified annotations for unknown metabolites. In contrast to existing methods, my approach that is based on metabolite levels, genetic associations, and information from public resources, sets unknown metabolites into their biochemical contexts (facilitating reliable interpretations), even if their identity cannot be resolved precisely in any case.

In this work I introduced two novel automatic approaches for biochemical clarification of unknown or ambiguously labeled metabolites that are especially based on measured metabolite levels. Both approaches enable or improve precise interpretations of scientific results in further metabolomics studies by identifying the local biochemical environment of metabolites and can be used alone or together with complementary methods.

Zusammenfassung

Experimente der Metabolomik dienen häufig der Erforschung von Gesundheit und Krankheiten, wobei Konzentrationen von Metaboliten als intermediäre Phänotypen erfasst werden, die verschiedene interagierende biochemische Ebenen vom Genom zu klinisch relevanten Phänotypen überbrücken. Vorab selektierte Metabolite können zum Testen von Hypothesen durch Technologien der *targeted* Massenspektrometrie (MS) gemessen werden. Das Gegenstück besteht aus der *non-targeted* MS, die alle Metabolite detektiert, welche sich in einer Probe befinden, und damit hypothesenfreie Tests zum Finden von unerwarteten Assoziationen ermöglicht, wobei sich darunter auch Metabolite mit zunächst unbekannter molekularer Struktur befinden. Während unbekannte Metabolite ohne jegliche strukturelle Annotation nur bei der *non-targeted* MS vorkommen, kann auch *targeted* MS zu mehrdeutig benannten Metaboliten führen. Der Wert von Experimenten der Metabolomik hängt stark von den gemessenen Daten und deren Annotation ab: Funktionale Metabolite oder Zwischenprodukte des Stoffwechsels beeinflussen die Messung von hochsensitiven Massenspektrometern, die sich in ihrer Massenauflösung (bspw. ± 0.3 oder ± 0.01) und in ihrer Leistung zur Trennung verschiedener Metabolite unterscheiden. Obwohl die Auftrennung verschiedener Metabolite durch chemische (bspw. Chromatographie) oder analytische (bspw. durch gezielte Selektion und Fragmentierung von Ionen in Tandem-MS) Methoden verbessert werden kann, können manche Metabolite, die ähnliche chemische und physikalische Eigenschaften besitzen, nicht unterschieden werden (z.B. Hexosen, Steroide, oder komplexe Lipide). Gemessene Metabolit-Summen tragen folglich mehrdeutige Bezeichnungen. Unbekannte oder mehrdeutig annotierte Metabolite in Ergebnissen von wissenschaftlichen Studien der Metabolomik stellen Sackgassen dar und können weder in biochemischen Zusammenhängen interpretiert, noch für klinische Anwendungen eingesetzt werden. Aus diesem Grund müssen unbekannte oder mehrdeutig annotierte Metabolite präzise charakterisiert werden. Einzelne unbekannte Metabolite können mit experimentellen Methoden, die u.a. auf MS beruhen, identifiziert werden, was allerdings mit sehr hohem Aufwand verbunden ist. Um Metabolite im großen Maßstab charakterisieren zu können, werden in den letzten Jahren computergestützte Hochdurchsatzmethoden bereitgestellt, die überwiegend auf Fragment-Spektren beruhen. Da einige kommerzielle MS-Plattformen Spektren nicht zur Verfügung stellen und Weiterentwicklungen an Massenspektrometern zu schnelleren Messungen, größeren Datensätzen, und zu einer steigenden Anzahl an unbekannten Metaboliten führen, wird die Metabolit-Charakterisierung auch in Zukunft einen wichtigen Stellenwert in der Metabolomik einnehmen.

In dieser Arbeit führe ich zwei neue automatisierte Methoden ein, die weitestgehend unabhängig von Fragment-Spektren arbeiten: (*i*) Bestimmung der hauptsächlichen Bestandteile von gemessenen Metabolit-Summen und (*ii*) Charakterisierung von unbekannten Metaboliten einschließlich der Selektion von Kandidatenmolekülen. Im ersten Teil stelle ich eine automatische Methode vor, die anhand eines Plattformvergleichs in einem kleinen Datensatz die Hauptkomponenten von Metaboliten mit mehrdeutigen Bezeichnungen bestimmt, die zum Imputieren von Konzentrationen eines (unabhängigen) großen Datensatzes auf einer höheren Präzisionsstufe genutzt werden können: Phosphatidylcholine (PCs), die aus einer polaren Kopfgruppe und zwei unpolaren Fettsäureketten bestehen, werden von den meisten etablierten Lipidomik-Plattformen gemessen, indem die Gesamtlänge der Fettsäureketten und deren Sättigung bestimmt wird, wobei die Aufteilung in die *sn*-1 und *sn*-2 Positionen nicht unterschieden werden kann (‘Lipidspezies-Level’). Moderne Plattformen messen Metaboliten auf einem höheren Präzisionslevel, indem sie die Länge und Sättigung beider Seitenketten getrennt bestimmen, wobei sie deren Reihenfolge und die Positionen der Doppelbindungen nicht angeben können (‘Fettsäure-Level’). Für dieses Experiment habe ich Konzentrationen von 76 PCs des ‘Lipidspezies-Level’ und 109 PCs des ‘Fettsäure-Level’ verwendet, die von Absolute*IDQ*$^{\text{TM}}$ (welche von vielen großen Kohortenstudien verwendet wurde) bzw. von Lipidyzer$^{\text{TM}}$ (welche präziser quantifiziert) in den selben 223 Proben von gesunden jungen Männern gemessen wurden. Zunächst habe ich 150 bis 456 theoretisch mögliche Bestandteile für die 76 gemessenen PC-Summen (‘Lipidspezies-Level’) bestimmt, indem ich die Länge und Anzahl an Doppelbindungen systematisch auf ihre *sn*-1 und *sn*-2 Positionen verteilt habe. 69 der 109 PCs wurden als Bestandteil von 38 PC-Summen identifiziert, wobei 10 Kompositionen aus einer hauptsächlichen Komponente mit einem quantitativen Beitrag von 80% bestanden. Weitere 17 Kompositionen enthielten ein oder zwei Komponenten mit einem Beitrag von 20 bis 80%. In einem Fall wurde ein PC als Hauptkomponente

ermittelt, welches eine ungerade Kettenlänge bei einer seiner Fettsäureketten aufwies. Diese Beobachtung war nicht zu erwarten, da geradzahlige Fettsäuren in tierischen Zellen bzw. Plasma wesentlich häufiger vorkommen als ungerade Kettenlängen. Da 13 quantitative Kompositionen direkt und 19 Kompositionen durch Imputieren von Metaboliten-Konzentrationen in Proben einer zweiten unabhängigen Kohorte repliziert wurden, kann davon ausgegangen werden, dass sich die bestimmten Faktoren zum Imputieren von PC-Konzentrationen eignen ('Lipidspezies-Level' → 'Fettsäure-Level'). Obwohl PCs des 'Fettsäure-Levels' nicht mit molekularen Strukturen gleichzusetzen sind, erlauben sie wesentlich präzisere Interpretationen im biochemischen Zusammenhang.

Im zweiten Teil stelle ich ein systembiologisches Modell vor, das die Charakterisierung von unbekannten Metaboliten, die durch *non-targeted* MS Technologien gemessen wurden, aus einer ganzheitlichen Perspektive angeht. Während sich bestehende Methoden mit dem Verknüpfen von aus Fragment-Spektren gewonnenen Informationen und Datenbankeinträgen beschäftigen, identifiziert mein Ansatz die biochemische Umgebung von unbekannten Metaboliten, bevor konkrete Kandidaten selektiert werden. Mit diesem Ziel habe ich 637 Konzentrationen (388 bekannt, 249 unbekannt) verwendet, die in 2 279 Proben gemessen wurden, um paarweise partielle Korrelationen zu bestimmen, die nachweislich ohne Vorwissen biochemische Stoffwechselwege rekonstruieren können. Das Modul zur automatischen Datenintegration vereinigt signifikante Assoziationen mit einem publizierten Gaußschen graphischen Modell (GGM), die mit 1 040 Kanten das Rückgrat des Netzwerkmodells darstellen. Des Weiteren wurden 299 Metabolit-Gen-Assoziationen einer publizierten mGWAS und 1 152 Reaktionen sowie 495 funktionale Metabolit-Gen-Relationen einer öffentlichen Datenbank zum Modell hinzugefügt, indem bereits integrierte Metabolite gefunden und weitere Metabolite ergänzt wurden. Zur Charakterisierung von unbekannten Metaboliten greifen mehrere Module auf das Netzwerkmodell zu und transferieren Informationen von bekannten zu unbekannten Metaboliten. Auf diesem Weg wurden 180 unbekannte Metabolite mit einer Zuordnung zu einem Stoffwechselweg annotiert und für 100 Paare von bekannten und unbekannten Metaboliten 21 Reaktionstypen selektiert. 12 zugeordnete Reaktionen wurden mit einem zweiten Kriterium (Retentionszeit) untermauert. In 1 209 Proben einer zweiten unabhängigen Kohorte, in der Metaboliten mit einer hochauflösenden *non-targeted* MS Plattform gemessen wurden, replizierte ich 18 von 26 Annotationen der Stoffwechselwege, die in beiden Datensätzen unabhängig voneinander vorhergesagt wurden. Für 139 der 315 hochauflösend

gemessenen unbekannten Metabolite, die im Netzwerkmodell bekannte Metabolite unter ihren Nachbarn haben, wurden 676 Kandidaten automatisch selektiert, indem öffentliche Datenbanken auf strukturell ähnliche Moleküle (zu benachbarten Metaboliten im Netzwerkmodell) mit gleicher Masse durchsucht wurden. Zum Schluss habe ich zwei von fünf selektierte, kommerziell erhältliche Kandidaten durch MS-Messungen experimentell verifiziert und eine Datenbankstruktur vorgestellt, die es ermöglicht, vorhergesagte oder verifizierte Annotationen von unbekannten Metaboliten zu organisieren. Im Gegensatz zu bestehenden Methoden platziert mein Ansatz, der auf Metaboliten-Konzentrationen, genetischen Assoziationen, und Informationen von öffentlichen Datenbanken beruht, unbekannte Metabolite in ihren biochemischen Kontext, was auch für Metabolite deren Struktur nicht gänzlich aufgeklärt werden kann, fundierte Interpretationen erlaubt.

In dieser Arbeit habe ich zwei neue automatische Methoden zur biochemischen Aufklärung von unbekannten bzw. mehrdeutig annotierten Metaboliten vorgestellt, die in erster Linie auf gemessenen Metaboliten-Konzentrationen beruhen. Beide Methoden können eigenständig, oder in Kombination mit komplementären Methoden verwendet werden und ermöglichen bzw. verbessern präzise Interpretationen in künftigen Studien der Metabolomik, indem sie das biochemische Umfeld von Metaboliten identifizieren.

Scientific contributions

In course of this thesis I contributed to the scientific community by publishing in peer reviewed journals and presenting at conferences or summer schools as listed below.

- Quell JD, Römisch-Margl W, Haid M, Krumsiek J, Skurk T, Adamski J, Hauner H, Mohney R, Daniel H, Suhre K, Kastenmüller G. **A platform comparison to characterize compositions of phosphatidylcholines measured as lipid sums in human plasma**. *Prepared manuscript, concurrent submission to the Journal of Lipid Research scheduled for 12/2018.*

- Quell JD. **Characterizing compositions of phosphatidylcholines measured as lipid-sums in human plasma – a platform comparison**. Conference talk. *8th Grainau Workshop of Genetic Epidemiology "From Association to Function"*. Grainau, D, 2018.

- Quell JD, Römisch-Margl W, Colombo M, Krumsiek J, Evans AM, Mohney R, Salomaa V, de Faire U, Groop LC, Agakov F, Looker HC, McKeigue P, Colhoun HM, Kastenmüller G. **Automated pathway and reaction prediction facilitates in silico identification of unknown metabolites in human cohort studies**. *Journal of Chromatography B – Analytical Technologies in the Biomedical and Life Sciences*, 1071:58–67, 2017.

- Quell JD. **What can we learn from population metabolomics?**. Conference talk. *1st Munich Metabolomics Meeting*, Freising, D, 2017.

- Hastreiter M, Jeske T, Hoser J, Kluge M, Ahomaa K, Friedl MS, Kopetzky SJ, Quell JD, Mewes HW, Küffner R. **KNIME4NGS: a comprehensive toolbox for Next Generation Sequencing analysis**. *Bioinformatics*, 33(10): 1565–1567, 2017.

- Altmaier E, Menni C, Heier M, Meisinger C, Thorand B, Quell JD, Kobl M, Römisch-Margl W, Valdes AM, Mangino M, Waldenberger M, Strauch K, Illig T, Adamski J, Spector T, Gieger C, Suhre K, Kastenmüller G. **The Pharmacogenetic Footprint of ACE Inhibition: A Population-Based Metabolomics Study**. *PLoS One*, 11(4):e0153163, 2016.

- Quell JD, Römisch-Margl W, Colombo M, Krumsiek J, Looker H, Agakov F, McKeigue P, Gieger C, Adamski J, Suhre K, Colhorn H, Kastenmüller G. **Identification of unknown metabolites with systems biology models**. Poster presentation. *SUMMIT Symposium and Plenary Meeting*, Malmø, Sweden, 2015.

- Quell JD, Krumsiek J, Gieger C, Adamski J, Suhre K, Kastenmüller G. **Identification of Unknown Metabolites with Means of Systems Biology**. Talk and poster presentation. *International Summer School: Data Aquisition and Analysis in Metabolomics*, Pula, Cagliari, Italy, 2014.

- Quell JD, Krumsiek J, Gieger C, Adamski J, Suhre K, Kastenmüller G. **Identification of Unknown Metabolites with Means of Systems Biology**. Poster presentation. *Conference Metabomeeting 2014*, London, GB, 2014.

Contents

CHAPTER 1

Introduction

In the last decades, several *omics*-disciplines have emerged in modern molecular biology. *Omics* is used as suffix that aggregates molecular analyses focusing on the entirety of specific molecular units of one kind. The prefix specifies the molecular layer: e.g. proteomics, genomics, lipidomics, or transcriptomics. Metabolomics, as an example, deals with questions about small molecules, such as lipids, amino acids, nucleotides, peptides, and carbohydrates [1]. Metabolites can be found in any cell or fluid of living organisms (kingdoms: prokaryotes, eukaryotes (animals, plants, etc.), or archaea). While specific metabolites are substrates, intermediates, or products of biochemical reactions or pathways, the metabolome describes all metabolites as entirety. [2, 3]

Subdivision of molecular constituents of biochemical processes into several *omics*-areas requires systems biology concepts (as super-structure) to reconstitute multi-level processes in integrative networks [4]. Although data in biology grows exponentially and is regularly organized as complex evolving networks of publications, topics, and ideas, there is still only a very tiny fraction of the theoretical information space (organisms × organs × cell types × genes × metabolites; and interactions, phenotypes, etc.) covered [5]. Efficient concepts and tools are necessary to utilize large parts of published knowledge for new projects. Systems biology tries to integrate various *omics*-modules to receive a comprehensive model containing *all* levels of biochemical function. This principle copes with the challenge to depict (parts of) the biological reality, in which many systems interact with each other. [3]

1.1 Current concepts and techniques in metabolomics

In contrast to genomes, the human metabolome is influenced by fixed (e.g. genetics, sex) as well as variable factors, such as the environment (air and water quality, diseases, lifestyle including specific diets, food, regular night shifts, and physical activity), molecular parameters (epigenetics, medication, and gene regulation), and personal properties (age and body mass index (BMI)). As a consequence, the metabolome can be seen as intermediate (molecular) phenotype that is (at least in part) objectively measurable in very large sample sizes [6, 7]. Some of these factors depend on each other (e.g. food and BMI, or disease and medication) and further ones can be both, cause or consequence of metabolic changes (e.g. diseases, medication, epigenetics, gene regulation, and BMI) [8]. The variability of metabolomes enables estimation of metabolic responses to environmental factors (quantitatively or qualitatively) and exploration of (functional) links between metabolite levels and genetic predispositions [9–11].

Metabolomics experiments are based on small clinical case-control or interventional study designs or on large cohort studies [9, 11]. In general, cohort studies employ an epidemiological observation study design in which several hundred or thousand participants reflect a cross-section through a certain population [12]. Participants are usually selected according to specific criteria, such as disease state, residential region, sex, age, BMI, medication, and ethnic origin. Appropriate protocols for collection of information and bio-samples (e.g. blood plasma, serum, urine, saliva, sweat, tears, or cell tissue) ensure unbiased data for later analyses. Besides metabolite levels, cohort studies incorporate further data on participants: e.g. clinical or life style information. These data are frequently collected by questionnaires that are very cheap, but the data quality strongly depends on the honesty, knowledge, and self-assessment of participants [13–15]. Because of cost, effort, and ethnics some data or samples are only collected for a small subset of cohorts. As an example, Cooperative Health Research in the Augsburg Region (KORA) (n = 3 080) focuses the exploration of cardio vascular diseases including cross-links to environmental factors and samples were randomly selected in the region of Augsburg [16, 17]. The German chronic kidney disease (GCKD) study (n ≈ 5 000) examines patients suffering from the homonymous disease [18], surrogate markers for micro- and macro-vascular hard endpoints for innovative Diabetes tools (SUMMIT) was sampled based on seven existing cohorts (n = 2 279) [19–23], and the German national cohort (NAKO) (n ≈ 200 000) will be a large data basis for investigating the functional origin of common diseases (e.g. cancer, diabetes, or cardio vascular diseases) and for finding new diagnostic and treatment procedures [24]. Since measured metabolomes represent snapshots of evolving systems, longitudinal studies, such as the Human Metabolome (HuMet) study [25], measure metabolite levels in samples that were taken during several time points.

Data from large cohort studies is often analyzed by metabolite-based genome-wide association studies (mGWAS), in which metabolite levels are statistically linked to genetics, i.e. to single nucleotide polymorphisms (SNPs), by linear models: genotype AA (major allele homozygous), Aa (heterozygous), or aa (minor allele homozygous). These analyses provide insights into probable tasks of gene products in biochemical pathways [26–30]. In metabolome-wide association studies (MWAS), (patho-) phenotypes are correlated to metabolite levels for identification of possible disease pathways [31–33].

MWAS and mGWAS are statistical approaches that do not provide information about causal relationships. Therefore, resulting links must be examined manually, using previous biochemical knowledge. There are many publicly accessible databases organizing results of statistical analyses and in part functional knowledge: e.g. GWAS server [34], SNiPA [35], ENCODE [36, 37], or Kyoto Encyclopedia of Genes and Genomes (KEGG) [38, 39].

1.1.1 Techniques to quantify metabolites

There are two fundamentally different high-throughput technologies to quantify large sets of metabolites: mass spectrometry (MS) (optionally a chromatography is coupled to the MS device) and nuclear magnetic resonance (NMR) spectroscopy. In this work, the focus is on data measured by MS, which is described in this section. Mass spectrometers consist of three basic parts: an ion source evaporates and ionizes molecules, an attached mass analyzer accelerates and filters ions according to their mass-to-charge ratio, before a detector determines the amount of passed ions.

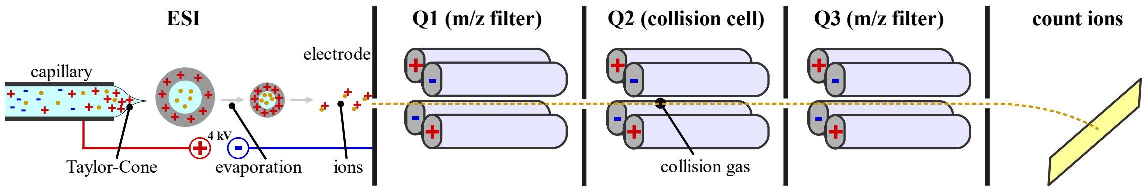
Sample preparation

Before quantification, metabolites need to be extracted according to established protocols that depend on the matrix and measuring technique. In general, bio-samples are homogenized, dissolved in solvents (e.g. methanol), and centrifuged to separate metabolites from proteins and cell organelles. Some metabolites may require further treatment for later ionization. Sample preparations should be simple and fast to ensure reproducibility and as non-selective as possible to keep a comprehensive set of metabolites. A careful sample preparation is essential, because metabolites that get lost during extraction cannot be restored later. [40–42]

Most metabolites remain stable over several years, if raw (or extracted) samples are stored at -80°C or in liquid nitrogen [43].

Mass spectrometric measurement

During the first step, an ion source accelerates and forwards molecules into the mass analyzer. Soft ionization methods, like matrix-assisted laser desorption/ ionization (MALDI) that is most efficient for peptides, or electrospray ionization (ESI) that is optimal for most of the other polar molecules and is a quasi-standard for metabolomics MS, lead to very few fragmentation of metabolites. An ESI source consists of a metal capillary (inner diameter around 100 µm), with a voltage applied to its end and an electrode (cathode in positive mode or anode in negative mode) containing a small hole aligned to the capillary (Figure 1.1(a)). Metabolites dissolved in methanol and water are lead through the capillary (flow rate $1-10$ µl/min). The high voltage of $3-4$ kV leads to an accumulation of positively (in positive mode) or negatively (in negative mode) charged ions leaving the end of the capillary through a formation called Taylor-Cone. The solvent of the released aerosol progressively evaporates until



(a) ion source (b) mass analyzer (triple quadrupole) (c) detector

Figure 1.1: Functional elements of a tandem mass spectrometer
Main elements of a tandem mass spectrometer are shown as exemplary sketch. (a) ESI generates ions based on molecules and a solvent. For the negative mode, lots of the capillary and the electrode are swapped. (b) The first and the third quadrupole (Q1 and Q3) filter ions according to their mass-to-charge ratio. The second quadrupole (Q2) is filled with a collision gas and fragments ions. The charge of electrodes changes, whereas opposite pairs of electrodes are applied with the equal charge. (c) Finally, the detector counts ions that pass through the mass analyzer. These parts of mass spectrometers are inside vacuum chambers. The elements of the functional sketch are not true to scale. Contents of this figure has been inspired by Grebe et al. 2011 [44] and Honour 2003 [45].

droplets burst and a stream of ionized molecules passes the hole of the electrode and enters the mass analyzer. [46–49]

Mass analyzers (e.g. time of flight (TOF), quadrupole, or ion traps) accelerate ions, optionally fragment metabolites, and filter them according to their mass-to-charge ratio. TOF analyzers accelerate ions and measure the time that ions need to move through the analyzer until reaching the detector. The velocity of equally charged ions only depends on their mass (ions of peptides may be multiply charged) [50, 51]. Quadrupole mass analyzer contain four rod-shaped electrodes, in which the same charge is applied to opposite electrodes (Figure 1.1(b)). An oscillating electric field moves ions on a circular path through the mass analyzer. The radius depends on their mass-to-charge ratio and on the energy of the applied electric field. Only ions with a mass-to-charge ratio within a predefined range can pass the quadrupole [52–54]. Ion traps (e.g. Orbitrap) collect ions in electrostatic fields and release them in a controlled manner by electric impulses or measure them inside based on their specific oscillation [55, 56].

Ions passing the mass analyzer are forwarded to a detector (Figure 1.1(c)). Detectors basically count the number of ions, leading to a semi-quantitative estimation of metabolite levels. The measurement of internal standards with known concentrations increase the precision that may (partially) allow an absolute quantification (within a device-dependent error range). [57, 58]

Separation of metabolites

Samples can either be directly injected into mass spectrometers (flow injection analysis (FIA)), or they went through a chromatography for a chemical pre-separation of molecules first. In the best case, MS quantifies preferably individual chemical structures. The procedure described above would lead to an excessive quantification of mixtures of several compounds with a narrow mass-to-charge range, because the separation of molecules is not sufficient. There are several approaches of tandem mass spectrometry (MS/MS) to select metabolites more precisely: e.g. using a triple quadrupole or an ion-trap mass analyzer. Furthermore, injecting probes through a chromatography before mass spectrometric analyses enable a pre-separation of analytes. Often with triple quadrupole mass spectrometers, sample extracts are directly injected into the ionizer (FIA). Mass analyzers of these mass spectrometers contain three quadrupoles that are connected in linear series. In multiple reaction monitoring (MRM) mode, ions are preselected by the first quadrupole (Q1) based on their mass-to-charge ratio. Passing ions get into the second quadrupole (Q2) that functions as collision cell. The cell contains a collision gas (e.g. nitrogen), whose molecules collide with ionized molecules that break into smaller parts. Resulting fragments are finally filtered by the third quadrupole (Q3) and are passed forward to the detector. An ion-trap (e.g. orbitrap) can combine several procedures in a single mass analyzer. Ions are collected within the orbitrap and specific ions are isolated by removing all other ions out of the analyzer. While triple quadruples perform a (*i*) selection, (*ii*) fragmentation, and (*iii*) selection step, orbitraps allow several iterations of selection and fragmentation runs. Precursor scans (MS^1 spectra) that are recorded before (without) fragmentation contain intensities of mass-to-charge ratios of complete ions. Selected mass-to-charge ratios (within a selected range and above a chosen intensity) are fragmented and recorded as MS^2 spectra. This spectrum contains fragmentation patterns of single peaks of the MS^1 spectrum. [59–61]

As second option to increase the separation of measured compounds, the mass spectrometer can be coupled to a liquid chromatography (LC) or gas chromatography (GC). Probes are injected into the chromatography column, leading to a pre-sorting of metabolites according to their retention time (RT), before they are (physically separated) forwarded to the attached mass spectrometer. Therefore, two different techniques (chromatography and mass spectrometry) targeting different chemical or physical properties are used for distinguishing metabolites. In general, chromatographic procedures use a stationary phase that retains target molecules and a mobile phase that washes out molecules. Molecules that fulfill the following three criteria can be physically separated by this procedure: they need to be retained by the stationary phase, are solvable in the mobile phase, and show different affinities to the stationary phase (to be washed out successively). The column is developed with a gradient of the mobile phase solvent (increasing concentration) during a defined period. The time molecules need to pass the chromatography column is measured (RT) and represented by a total ion chromatogram (TIC) (x-axis: retention time, y-axis: base peak intensity) or for a selected mass-to-charge ratio (i.e. specific molecule) by an extracted ion chromatogram (EIC). RT is usually standardized receiving a retention index (RI) to increase comparability. The columns of LCs consist of a capillary with a connected high-pressure pump for probe injection. The mobile phase usually contains organic solvents (e.g. methanol). [62, 63]

Targeted and non-targeted metabolomics

Measuring techniques of metabolomics can be generally classified as targeted or non-targeted. In targeted metabolomics a predefined set of metabolites of interest (specific group or metabolites of one pathway) is measured with high accuracy (absolute quantification in nmol/g or Micromolar ($\widehat{=}$ µmol/L)). The extraction can be specific for the selected compounds, leading to a selective extraction, higher level of purification, and better probe quality. Since all metabolites are (accurately) identified, this approach is capable for e.g. hypothesis testing. Measurements can be reproduced within a given error range (that is specific for the quantification technique). [64–67]

Non-targeted metabolomics (also known as metabolic fingerprinting) is a label-free approach for measuring as many metabolites that are in the sample (after preparation and extraction) as possible to capture the comprehensive cellular metabolism. The focus is on the detection of possibly thousands of metabolites of various groups and pathways. Samples require a comprehensive preparation (several runs applying different techniques) for extraction of a heterogeneous set of metabolites. In return, measured metabolites are previously not annotated by chemical structures and only a subset can be identified (depending on the MS platform). Non-targeted metabolomics typically leads to semi-quantitative measurements (ion counts), because no further knowledge and no internal standards are available. As another consequence, measurements cannot be easily reproduced and metabolite levels of different measurements (measuring time points, laboratories, or platforms) cannot be compared directly. This limitation is in part resolved by using metabolite ratios (ideally with a pair of metabolites that are causally related) or metabolite associations to stabilize the signal and therefore enable reproducibility and comparability. Non-targeted metabolomics is suitable for large scale screens (for obtaining patterns or comprehensive fingerprints) and hypothesis-free assays (e.g. MWAS for identification of new biomarkers). [66–70]

In general, quantified metabolites may consist of mixtures of compounds instead of single chemical structures (e.g. hexoses, or complex lipids). Some metabolites cannot be separated by MS (and optionally by coupled LC or GC) according to their chemical and physical properties. Quantification of metabolites on mass spectrometric devices is influenced by the extraction procedure, temperature, air pressure, humidity, the executing laboratory, equipment, and lots of consumables.

Depending on the weight of quantified molecules, the usable resolution of older MS devices is ± 0.3 Da and for newer devices $\pm 0.0001 - 0.001$ Da (accurate mass resolution). Molecular masses or mass differences are reported in the chemical standard units Dalton (Da) or unit (u). These units are defined as relative measure according to the atom mass of carbon-12: $1\ Da = 1\ u = \frac{mass\left({}^{12}C\right)}{12}$. Mass spectrometers usually measure the mass-to-charge ratio $(\frac{m}{z})$ that is equivalent to Da, if $z = 1$. The mass accuracy of high-resolution MS measurements is indicated in parts per million (ppm) and depends on the measured mass ($ppm = \frac{m_{measured} - m_{exact}}{m_{exact}} \cdot 10^6$) (e.g. $0.5 - 1$ ppm for an orbitrap mass analyzer). While the exact mass of molecules refers to a calculated mass of molecules (according to a specified isotope distribution), the accurate mass is defined as measured mass that enables determination of the unambiguous atomic composition of molecules ($\ll 0.1$ ppm, depending on the molecular weight). Public databases including Human Metabolome Database (HMDB) [71], KEGG [38], PubChem [72], and ChemSpider [73] report the monoisotopic mass (sum of masses of the most abundant isotopes of atoms of a molecule) and the average mass or molecular weight (sum of the average masses of atoms of a molecule according to an observed isotope distribution) for each molecule, considering different isotopic compositions. [74–77]

There are several standardized platforms (e.g. Biocrates$^{\text{TM}}$ [1] Absolute*IDQ*$^{\text{TM}}$ kit for targeted metabolomics or Metabolon$^{\text{TM}}$ [2] for non-targeted metabolomics) for quantification of a set of metabolites. These companies provide procedures for sample extraction, measurement, and analysis of samples, specify measuring devices and offer consumables (solvents, chemicals, chromatography columns, standards, etc.) and software. In part, they carry out metabolite identification and quality control (QC) of measured data. Metabolon$^{\text{TM}}$, as an example, identifies a subset of measured metabolites based on RI (in the EIC) and MS^2 spectra (fragmentation pattern) according to their internal library. Quantification is performed by the base peak intensity.

1.1.2 Challenges and opportunities of metabolomics experiments

Metabolomics has great potential in clinical applications, because measurements require very small amounts of probe material (usually 10 µL − 1 mL) and high numbers of (several hundred or thousand) metabolites can be quickly measured in easily accessible matrices like blood, urine, or saliva. Measured metabolomes reflect the physiological status that depends on intrinsic (gene variation or diseases) or extrinsic

[1] Biocrates$^{\text{TM}}$ Life Sciences AG, Innsbruck, Austria

[2] Metabolon$^{\text{TM}}$ Inc., Durham, NC, USA

(environmental) factors. Although metabolomics platforms can only capture a subset of the metabolome of an organism, quantification incorporates several hundred or thousand metabolites. Compared to traditional quantification of single values, metabolomics quantification is cheap and scalable to large sample sizes.

One of the major challenges of metabolomics (especially for non-targeted metabolomics) is the limited comparability of ‘concentrations’. However, it has been shown, that ratios of (functionally related) metabolite levels that have been measured in different laboratories or on different devices are consistent across methods and protocols [78, 79]. This principle is successfully applied in new born screens that are established in clinical context since several years [80, 81]. As second solution, standardization of metabolite levels increases comparability significantly.

Non-targeted metabolomics leads to measurement of distinguishable metabolites with an initially unknown chemical structure. The measuring device provides RT (if a LC or GC hast been performed), mass-to-charge ratio, and (depending on the executive laboratory or company) fragmentation spectra, but the chemical identity or classification into biochemical pathways are lacking. Results of scientific analyses (e.g. cohort studies) frequently contain associations to, regressions with, or models containing unknown metabolites. The further use of these results is limited, since interpretations in biochemical contexts or clinical applications are impossible. This procedure requires chemically unambiguously characterized, or at least partially annotated metabolites [82]. While previous metabolite characterization approaches compare complete measured and target fragmentation spectra [83–85], recent methods cluster individual peaks [86], perform *in silico* fragmentation of database structures for a comparison to measured spectra of unknown metabolites [87–89], predict molecular fingerprints based on measured spectra of unknown metabolites to select database compounds with similar fingerprints [90–93], or construct networks based on spectral similarity and known biochemical relations [94]. Characterization of unknown metabolites enables scientific interpretations by placing respective molecules into their biochemical context.

On a cellular level of biochemical pathways, some metabolites have very high turnover rates (assuming a normal distribution of turnover rates), leading to flexible metabolic fingerprints that strongly depend on the type or physiological state of matrices (or tissues) and compartmentation of pathways [82, 95]. Levels or measured concentrations of specific metabolites are influenced by various factors based on subjects (genetic variations, sex, age, BMI, lifestyle, food, diseases, drugs, and microbiome), preanalytical probe handling (collection, storage, and extraction), or measurement (batch effects and metabolomics platform) [96]. Although levels of single metabolites react sensitively to diseases, drug intake, or environmental exposures, cohort studies showed that individual metabolomes are stable over several years [6, 7]. Metabolomics measurements show singular snapshots of an evolving biochemical system. Longitudinal analysis (taking several samples within defined periods of times to model time series) can partially cope with this issue, however, models can impossibly depict the entirety of influencing factors. The sample size should be appropriate and major driving factors should be covered.

Metabolomics has an important role in current scientific health or disease related studies. This includes examination of influences of diseases, environmental factors, or genetic predispositions to the biochemical balance determined by mGWAS or MWAS as first analyses. It is still a challenge to transfer observations from systems biology models to clinic cases, because some impacting factors, such as diet or lifestyle, cannot be completely observed. Cross-sectional cohort studies usually contain large sample heterogeneity that is challenging for learning procedures, but also for procedures such as prediction or classification [95].

As hypothesis driven as well as hypothesis free approaches, metabolomics makes sense in clinical medicine for several applications: e.g. discovery of hidden reasons for common diseases may lead to disease-origin oriented treatments (in contrast to conventional treatments of symptoms solely) or investigation of individual rare diseases (cf. Chapter 1.2.3). It is an ongoing challenge to convince physicians of using metabolomics analyses in their everyday practice and health insurances to pay for these diagnostic examination techniques. Modern medicine is becoming more and more expensive (magnetic resonance imaging, surgeries, specialized medication, increasing number of treatment options) that will be a heavy burden for the social health system. Metabolomics can reduce the cost of laboratory blood tests dramatically and in addition, it provides several thousand values [97]. Systems biology implements strategies for integration of large datasets (e.g. high number of metabolites) to receive models that organize complex data and are easily searchable. [98]

1.2 Systems biology models support metabolomics studies

In many disciplines, traditional science is very successful by reducing tasks and answering questions isolated from their comprehensive environment. Since (micro) biological processes are typically highly crosslinked to other processes related to various regulatory levels, these systems can only be captured by a global description containing as much interacting factors as possible. Systems biology copes with the challenge to integrate various levels of *omics*-areas into a comprehensive model, knowing that (by definition) models depict only parts of reality. [99, 100]

Already fifty years ago, the central dogma of molecular biology dealt with the linear transfer of sequential genetic information and stated that this information cannot be transferred from protein to protein or back to nucleic acids [101]. However, there are processes influencing or regulating the gene expression (transcription, translation, and protein function) incorporating epigenetics/ histone modifications, microRNAs, cofactors, genetic predispositions, and environmental signals [102, 103]. In consequence, elements of any biological level can be both, source and target of regulatory processes.

In addition to the complexity of the matter, the amount of data in biology grows exponentially and leads to complicated evolving structures of papers containing ideas and information that cannot be handled manually. Models are required to make use of, apply, or combine this knowledge for generating a holistic picture of certain processes. In systems biology, data is integrated systematically, and hypotheses are tested on resulting computational or mathematical models. On the one hand, resulting models only depict a small subset of all elements of the biochemical reality. On the other hand, these models structure available information very well, are searchable, and easy to handle compared to various types of raw data [5].

1.2.1 Networks for organization of biological information

Systems biology provides concepts and tools to integrate large amounts of heterogeneous data of various *omics*-areas for approaching questions on comprehensive biochemical processes. To this end, relationships among interacting elements and their effect on observed parameters can be modeled, capturing the *big picture*. Possible elements include genes, mutations, proteins, metabolites, environmental parameters, age, sex, BMI, disease states, medication, nutrition, and life style habits. These models are not a complete copy of reality. Systems biology typically integrates large amounts of high dimensional and complex data (*big data*). Depending on measurement techniques and laboratories, the quality of data varies, missing values are produced, or the structure of information changes. [3, 95, 98, 99]

The underlying data-organization of most systems biology models consist of network-like relationships of elements. Nodes represent elements (e.g. genes, proteins, or metabolites) and edges represent interactions between elements (that either belong to the same level, e.g. protein-protein, or connect different levels, e.g. metabolite-enzyme). As an example, protein-protein interaction networks can incorporate several species and vast evolutionary distances [104] or metabolite-gene networks are used to reconstruct biochemical pathways including links to diseases, such as individual types of tumors, obesity, or diabetes [95, 105].

The distributions of edges and nodes of most networks in biology are comparable to scale-free topologies (Figure 1.2(b)) [106]. In contrast, random networks show topologies, in which all nodes have similar node degrees, which is very rare in biology, since self-organizing systems prefer established nodes, paths, or subsystems leading to scale-free topologies. Furthermore, processes or compartmentation that contain largely independent substructures can be modeled as hierarchical networks. Hierarchical networks (e.g. signaling models) contain several scale-free networks connected by hierarchical super-structures. In common scale-free network topologies, few metabolites are connected to many neighboring nodes (hubs, high degree) and the clear majority of nodes has a very small number of neighbors (low degree). As a consequence, these networks in biology hold the small-world property, meaning that the distance of any pair of nodes is relatively short (distance: shortest path between two nodes; path: number or weights of edges) [107]. Essential genes that encode housekeeping proteins or -pathways tend to be represented by hubs. Scale-free networks are robust against random failures, because probably nodes with small degrees are affected leading to a preservation of the main network structure and function. However, this type of networks is sensitive to 'intended attacks'. In the worst case, networks may break into several parts, if main structures (such as single hubs) are affected. In the best case, there are alternative routes through the network, but they may have reduced capacities. Even for large networks, there are algorithms for finding these potentially weak vulnerabilities efficiently. In clinical applications, hubs are promising starting points for biomarker- or drug target discovery. [106, 108, 109]

Most networks of biological information contain modules that are substructures showing a high degree of clustering (highly connected part of the network). Smaller substructures that are related to a specific role (e.g. (negative) feedback loops, (positive) feed forward loops, bifans, or oscillators) and frequently occur in biology networks are called motives. Network models are accessed by algorithms to detect certain modules, to search for shortest paths, to identify clusters, or to discover unexpected relations. Moreover, network models are able to combine results of several experiments to perform meta analyses or answer new questions [98].

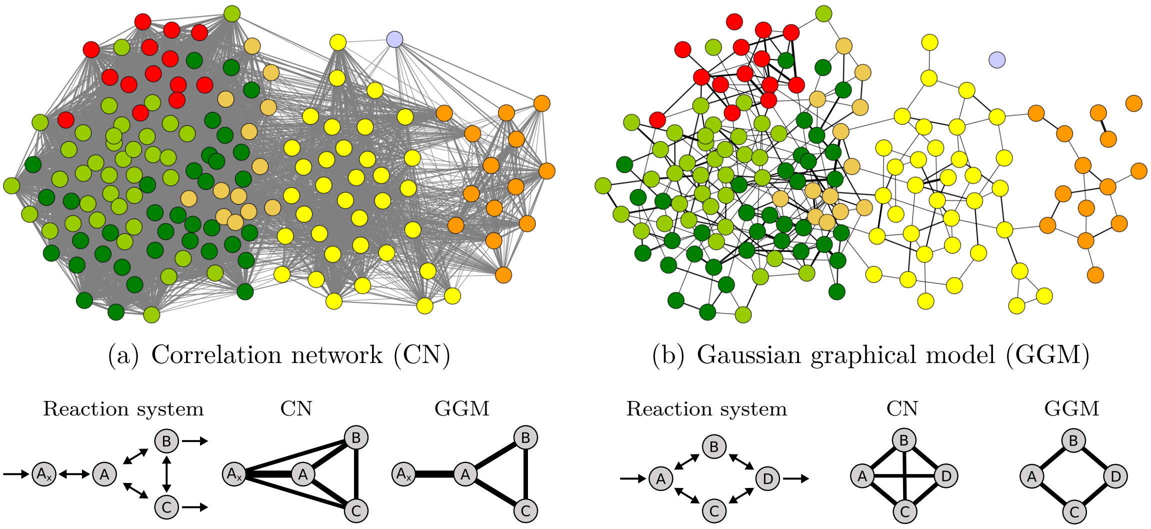
1.2.2 Applications of network models in metabolomics

Organizing highly connective metabolic pathways requires systems biology models. Metabolic networks contain information such as reactions between metabolites, pathway annotations, and genes that encode proteins catalyzing these reactions. Although, metabolic networks belong to the most comprehensively described networks in biology, it is still an ongoing challenge to reconstruct the human metabolome (Recon: Virtual Metabolic Human Database [110, 111]) as a whole. Elements of metabolic networks are clearly separated: e.g. metabolites, reactions, and enzymes. However, there are cofactors, intermediates, co-substrates, byproducts, or secondary compounds (e.g. ATP, ADP, NADPH, $NADP^+$, or Water) leading to large amounts of interconnections of metabolic networks. These compounds are necessary for, or produced by organisms, but play unspecific roles in many reactions. Furthermore, pathways that are interwoven with (or at least connected to) the basic housekeeping metabolism have an increasing effect on the number of distant crosslinks and each level of biochemical organization and control (genomics, gene expression, protein expression, and metabolism) is often measured in different timescales. Since metabolites are highly cross-linked and influenced by various stable and flexible factors, systems biology typically follows a top-down approach, in which observations of the general system are transferred to a special part of its sub-systems (deduction). [82, 112, 113]

Biochemical pathways can be reconstructed manually, hypothesis-driven, based on experiments in the wet-laboratory, or in smaller scales computationally. Resulting network maps are very precise in the basic (housekeeping) metabolism, but contain little information about uncommon pathways, such as special disease or environmental effect response pathways, or unexpected interconnections between pathways (e.g. Recon: Virtual Metabolic Human Database [110, 111], KEGG pathway maps [38, 39], MetaCyc/ HumanCyc/ BioCyc [114, 115], or Roche biochemical pathways [116]). These general pathway models can be refined to genome scale metabolic models, if they are enriched with species genomes.

In contrast to genome scale metabolic models, metabolomics analyses use kinetic models that consist of directed networks quantitatively depicting fluxes in biochemical reactions and pathways. Therefore, environmental influences and allosterically regulated enzymes (usually only a few enzymes are regulated) are modeled precisely based on *in vitro* experiments. Metabolite levels are adjusted to achieve a balanced flux through each enzyme of the pathway. Large numbers of parameters and necessity of experimental results lead to complex analyses. However, those models can simulate the behavior of metabolomes over time, depending on defined input parameters. Some ressources are publicly available (e.g. BRENDA: Comprehensive Enzyme Information System [117] contains organism information about enzyme kinetics and turnover rates). [82]

While the approaches described above are based on previous biochemical knowledge, data-driven approaches lead to hypothesis-free reconstructions of biochemical relations incorporating measured known and unknown metabolites. The systematic estimation of pairwise Pearson correlations of metabolite levels of cohort studies leads to correlation networks (CNs) with significant correlations between almost all metabolites (Figure 1.2(a)). Transitive edges or edges between co-regulated metabolites need to be removed for the creation of realistic biochemical pathways (Fi-



(a) Correlation network (CN)

(b) Gaussian graphical model (GGM)

(c) Exemplary coherence I

(d) Exemplary coherence II

Figure 1.2: Comparison of correlation networks and Gaussian graphical models
Pairwise (a) standard Pearson product-moment correlations and (b) partial correlations were estimated by 151 metabolite levels of a cohort study ($n = 1\,020$) to receive a correlation network (CN) and a Gaussian graphical model (GGM). Colors of nodes refer to metabolite classes: yellow: Acyl-carnitines, green: diacyl-phosphatidylcholines, beige: lyso-phosphatidylcholines, light green: acyl-alkyl-phosphatidylcholines, red: sphingomyelins, orange: amino acids. (a & b) Gaussian graphical models (GGMs) are known to arithmetically reconstruct biochemical reactions that can be clustered to pathways, while CN often contain all of the possible edges ending up to harry balls. Reprinted figure from Krumsiek et al. by permission from BMC Systems Biology [118], copyright 2011.

gure 1.2(c & d)). This goal is achieved by partial correlations that consist of Pearson correlation coefficients conditioned against the correlation with all remaining metabolites. The conditioning leads to an isolation of direct effects by removing linear effects of all other metabolites on the tested variables (removal of indirect associations). The union of all significant pairwise partial correlations results in an undirected Gaussian graphical model (GGM) that is much sparser than CNs (Figure 1.2(b)). It has been shown that this hypothesis-free approach is able to robustly reconstruct biochemical reactions and pathways by only using metabolite levels of a cross-sectional cohort study. Further data, previous knowledge, or assumptions are not necessary. As second advantage, GGMs incorporate metabolites with an unknown chemical identity and place them into the specific biochemical context. [118]

Metabolomics with respect to systems biology concepts applies methods including network inference, network analysis, enrichment analysis, pathway analysis, flux analysis, metabolic modeling, or kinetics [112]. In general, correlations or relations do not necessarily imply causality (correlation is necessary bot not a sufficient condition to identify causal relations). Some approaches incorporate any information of different biochemical organizational levels hypothesis-free, but e.g. flux analyses require assumptions indicating the direction of relations, or causal links are manually modeled using knowledge of underlying real molecular mechanisms. [82, 119]

1.2.3 Opportunities in systems medicine

Most human diseases come along with changes in the metabolome. It is often assumed that diseases causally lead to perturbations in metabolic processes, but the opposite direction, or no causal relationships would also be possible. Patterns that can be identified in plasma or urine samples show a mixture of these (potentially causal) links. Furthermore, there are genetic and environmental factors that have causal effects on the living organism. [82]

In traditional medicine, most approaches for diagnoses, prognoses, and treatment of diseases have been established by correlations between clinical parameters and patho-phenotypes (reductionist approach). Although this process has been successfully applied to numerous diseases and is accepted by classical medicine, it struggles with the identification of disease-relevant functional relations in complex systems. Future medicine will cope with this challenge and includes systems medicine, precision medicine, network medicine, and personalized medicine that are described in the following. In general, future medicine incorporates diagnoses of established common or rare diseases, defining disease predilection, developing (individualized) treatment strategies of disease causes and explore comorbidities. [120]

While rare diseases are typically caused by a mono genetic defect that leads to a (severe) (patho-)phenotype (and may express itself in several symptoms), common diseases (diabetes, obesity, asthma, etc.) are the result of perturbations of the whole complex of intra- and intercellular processes [107]. Challenges, but also strengths of systems medicine are handling large amounts of heterogeneous data (*big data*) and identification of causal biomarkers or drug targets. Especially the transfer of insights from systems biology models to clinical cases is error-prone, because impacts of diets, lifestyle or further factors have frequently not been observed completely [95].

Network medicine systematically integrates genetic, genomic, biochemical, cellular, physiological, and clinical data into networks (disease networks, phenotypic networks) to model disease expression and underlying processes. Besides the classical observation of late stage diseases, systems medicine explores pre-phenotypes enabling early treatments to stop the development or manifestation of diseases [120]. Systems biology models enable hypothesis-free approaches for receiving unbiased insights into molecular relations. Disease genes that often interact with each other can usually be located in the network periphery (otherwise perturbations would not be compatible with life). These disease modules help to identify disease pathways and tend to appear comorbidity [107, 121].

The diseaseome consists of a systematic collection of overlapping disease modules that have a common genetic basis, share metabolic pathways, or often occur simultaneously (comorbidity) [122]. The opposite is also true: diseases that are located in separated modules are usually phenotypically distinct. Metabolomes play a major role in systems medicine, because metabolic levels directly relate to the current biological status of an organism (in contrast to the genome that is relatively stable over life time). Time-dependent changes and effects of drugs or environment can be systematically studied [119].

Classical epidemiological cohort studies are focused on public health by exploration of MWAS, risk biomarkers, or hypothesis testing. Precision medicine and personalized medicine target functional biomarkers for stratification and for prognosis of disease development. In future, individualized drug therapies, life style, nutrition, or environmental exposure will be recommended according to the individual status. Most promising drugs and their dosage will be individually selected to maximize the therapeutic impact and to minimize side effects. Those analyses use insights of systems medicine and add individual profiling of the metabolome and clinical parameters. [95, 119, 123]

1.3 Approaching structurally uncharacterized metabolites

Metabolomics has established as an important discipline in research of health and diseases, whose scientific outcome of experiments strongly depends on the quality of measured metabolites. Non-targeted metabolomics based on LC-MS has emerged as an established technology to simultaneously measure the levels of a wide range of low weight molecules in biofluids and tissues [68]. While this hypothesis-free approach allows to explore unexpected relations, large fractions of quantified metabolites are structurally unknown [124, 125]. Even targeted metabolomics includes measurement of (in part) poorly characterized metabolites. Consequently, scientific results containing these metabolites can hardly be set into biochemical contexts and their usage for clinical applications is very limited.

1.3.1 Uncharacterized metabolites in scientific results

For enabling reliable biochemical interpretations to discover new biochemical relationships based on scientific results containing unknown or ambiguously characterized metabolites, precise annotations of these compounds are critical. The metabolite identification task group of the Metabolomics Society Inc. is in line with this assessment and accentuated the community consensus that identification of measured unknown metabolites in large scales is one of the most important challenges in current metabolomics research. There are four levels of confidence, describing the range between unknown compounds that are reproducibly detected and quantified but not further annotated (level 4) to identified metabolites (level 1, verified by measurements of pure substances). In addition to level 4-compounds, putatively characterized compounds (label 3) come along with spectral or chemical/ physical properties that are consistent with other metabolites of a certain class of compounds. Putatively annotated compounds (level 2) need to fulfil the criteria of level 3-compounds and moreover, spectral similarities to respective spectra of public or commercial libraries need to be demonstrated. Increasing the amount of well annotated metabolites (low level number) is aimed to maximize the usability of quantified metabolites. [126–128]

Indeed, some metabolites in targeted as well as in non-targeted metabolomics are characterized, but their labels indicate an ambiguous structural annotation. Sets of molecules instead of single compounds are measured, if the applied analytical method cannot separate compounds that share equal chemical or physical properties (e.g. hexoses, steroids, or complex lipids). As a consequence, metabolite labels express specific precision levels: phosphatidylcholines (PCs) that are among the major constituents of cell membranes [129, 130], important compounds of lipoproteins [131], and crucial in the energy metabolism, consist of a polar head group with phosphocholine and a nonpolar part with two fatty acid chains (with different lengths and desaturation) connected to a glycerol [132–134]. PCs are frequently measured during MRM scans by selecting the total m/z (Q1) and the head group (Q3). As a consequence, the total length of both fatty chains and number of double bonds is determined, but their division and configuration at the *sn*-1 and *sn*-2 positions remain covered (lipid species level, cf. Figure 3.1). Furthermore, PC and isobaric PC-O, in which one fatty acid chain is longer by one and attached by an alkyl bond, are not analytically separated. However, slight changes in the chemical structure of PCs imply large functional differences, such as the biochemical usage, the membrane fluidity, or associations to diseases [135–137]. Therefore, meaningful biochemical interpretations need to be as close as possible to the chemical structure of lipids. That is currently only possible under assumptions or concrete expectations concerning their constituents [138–141]. Today, the availability of modern high-throughput lipidomics platforms allows quantification of numerous lipids in thousands of blood samples of epidemiological cohorts. In particular, the MS-based targeted quantification of metabolites, such as PCs, lysophosphatidylcholines (lysoPCs), or sphingomyelins (SMs) of the lipid species level, using Absolute*IDQ*$^{\text{TM}}$ has produced large data sets. In various large-scale mGWAS and MWAS, multiple of these PC measures have been reported to associate with common genetic variants and various diseases such as Alzheimer's disease [142, 143], Type 2 Diabetes [141], coronary artery disease [144], or gastric cancer [145]. The gap between the lipid species level and the specific chemical structure needs to be reduced to use results based on these measured metabolites reasonably.

1.3.2 Current metabolite annotation approaches

Singular unknown or ambiguously characterized metabolites are successfully annotated by traditional methods in wet laboratories. However, these approaches are very expensive and time-consuming, because they are limited to individual compounds, require numerous measurements, and findings concerning one compound cannot be transferred to the identification of other molecules. To enable metabolite annotations at large scales, research on *in silico* approaches was started a few years ago and has nowadays emerged to a hot topic in metabolomics [146].

Most *in silico* approaches for identification of unknown metabolites primarily focus on fragmentation spectra and only require measurement of one single sample. The previous methods compare complete measured fragmentation spectra to target spectra of public databases for selection of candidate molecules [83–85]. To bridge varying availability and comparability of fragmentation spectra, MetAssign [86] clusters m/z and intensity of all peaks of measured and target spectra separately and ranks possible candidates. Further approaches use *in silico* fragmentation of database structures to become independent from spectral databases: As an example, CFM (competitive fragmentation modeling) [87] uses a Markov-process to subsequently predict fragmentation events (fragmentation tree) including resulting fragments for chemical structures, or ranks possible structures of public databases based on measured MS/MS spectra. MetFrag [88] applies *in silico* fragmentation to database structures and calculates their expected RTs to substantiate candidates for unknown metabolites. The compMS^{2}Miner [89] benefits from combining several approaches including *in silico* reconstruction of fragmentation, noise filtration, substructure annotation by similar fragmentation and database spectra, functional group detection, and nearest neighbor chemical similarity scoring. A further class of approaches that work independent from spectral databases predict molecular fingerprints (e.g. PubChem fingerprint) based on fragmentation spectra of unknown metabolites (in part supported by fragmentation trees) and query structure databases for compounds with similar fingerprints (in which similarity is estimated by the Tanimoto coefficient): FingerID [90], CSI:FingerID [91, 92], and SIMPLE [93]. While metabolite characterization approaches, described above, focus on structural similarity of unknown metabolites (determined from fragmentation spectra) to database compounds, MetMapR is a graph-based tool that estimates similarity based on fragmentation spectra and database information, such as enzymatic transformations or metabolite structural similarity [94].

1.3.3 Objective

In this work, I provide two automated *in silico* approaches for (*i*) clarification of ambiguously measured metabolites (esp. PCs) by a systematic platform comparison of targeted MS and for (*ii*) an automated annotation of unknown metabolites that have been measured by non-targeted metabolomics including selection of specific candidate molecules by a systems biology model. Unlike existing approaches, both concepts are independent from fragmentation spectra (since spectra are frequently not available if commercial measuring kits or platforms are used), highly automated, and integrate primary data of measurements within cohort studies (e.g. concentrations or ion counts, mass-to-charge ratios, and RIs) or database information.

In the first part of this work, I characterize the composition of 76 PC-sums (lipid species level) measured by Absolute*IDQ*™ in 223 human plasma samples of healthy subjects by a platform comparison to 109 PCs (fatty acid (acyl/alkyl) level) measured on the Lipidyzer™[3] platform (Chapter 3). Measurements on the fatty acid level include determination of the lengths and desaturation of both fatty acids by MRM scans that select the total m/z (Q1) and one of both fatty acid chain (Q3). Their ordering (*sn*-1 or *sn*-2), position of double bonds or stereo-chemistry is not detected. An automated procedure estimates respective fractions of constituents and tested their variation during a replication in samples of a second independent cohort. Described quantities of constituents of PC-sums enable advanced interpretations of large amounts of scientific results that are based on measurements on the lipid species level and furthermore allow imputing concentrations receiving PC concentrations measured on the fatty acid level. Although the fatty acid level is far from molecular structures, imputation enables advanced interpretations of scientific results by clearly reducing the number of possible constituents.

The second part of this work targets metabolites, whose structures are completely unknown. Six years ago, Krumsiek et al. combined partial correlations between pairwise metabolites (GGM) and metabolite-gene associations (mGWAS) facilitating manual annotations of unknown metabolites [147]. In this work, I considerably extended and transferred this idea to fully automated *easy to use* workflows (Figure 1.3). Comprehensive data integration of GGM, mGWAS, and reactions and metabolite-gene relations of public databases (such as Recon [110]) provides a flexible data basis (network model) for following automated analyses (Chapter 4.2). I predict reactions connecting known and unknown metabolites based on measured mass-to-charge ratios as previously proposed by Breitling et al. [148] and predict pathways to identify the biochemical contexts of unknown metabolites (Chapter 4.3). In the end, I introduce an approach for automated selection of selected candidate molecules for unknown metabolites based on structural similarities to database compounds (Chapter 4.5)

[3] AB Sciex™ Pte. Ltd., Framingham, USA and Metabolon™ Inc., Durham, NC, USA

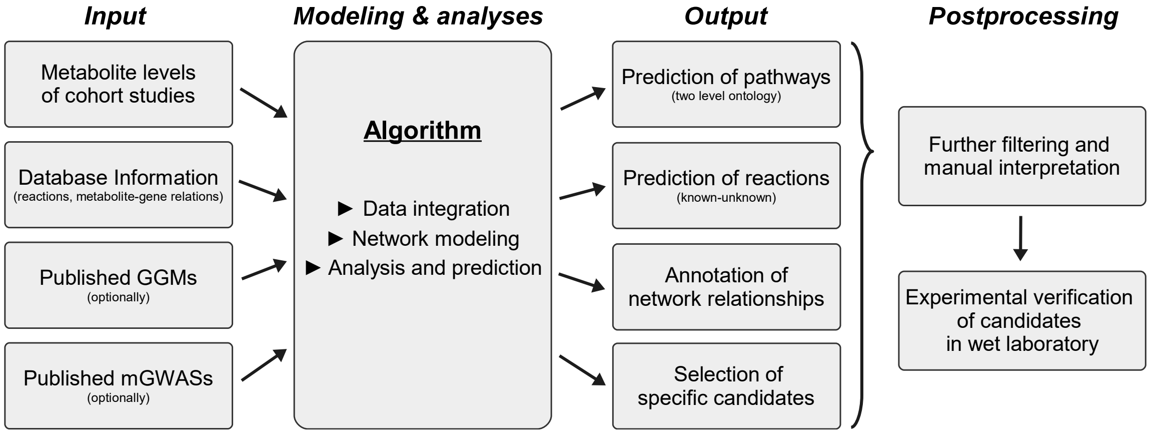

Input
Modeling & analyses
Output
Postprocessing
Metabolite levels of cohort studies
Database Information
(reactions, metabolite-gene relations)
Published GGMs
(optionally)
Published mGWASs
(optionally)
Algorithm
► Data integration
► Network modeling
► Analysis and prediction
Prediction of pathways
(two level ontology)
Prediction of reactions
(known-unknown)
Annotation of network relationships
Selection of specific candidates
Further filtering and manual interpretation
Experimental verification of candidates in wet laboratory

Figure 1.3: Process for characterization of unknown metabolites
The automated process starts with modules for input of data, performs modeling and analyses, and finishes with several outputs. Adaption of data to the input-interfaces and postprocessing are the only steps of this concept that consist of manual work.

and generate easily readable summaries of any information collected per unknown metabolite. Moreover, ideas for storage of predicted and verified information about unknown metabolites are proposed (Chapter 4.7). The automation of data integration and standardized analyses are the basis and key part of this work. Manual steps are systematically reduced to a minimum and basically consist of adaption of new input data to the provided interfaces and consistency checks of the output. These workflows are publicly available and include standardized interfaces enabling integration of different types of data. To demonstrate the applicability of the introduced workflow for metabolite identification, I applied my approach to a non-targeted metabolomics data set (439 known, 319 unknown metabolites) measured in blood samples of 2 279 subjects that were analyzed in the course of the project SUMMIT using a commercial LC-MS-based metabolomics platform (Metabolon$^{\text{TM}}$). To enable an automated selection of specific candidate molecules, I applied my approach to 1 456 metabolites (796 known, 660 unknown) of a second data set that has been measured by a non-targeted high-resolution LC-MS platform in 1 209 samples of the GCKD cohort. For a selected group of predicted metabolites, for which the pure compounds were commercially available, I tested candidate molecules experimentally.

The focus of both concepts is on the provision of tools and concepts to facilitate the further usage of scientific results containing unknown or ambiguously characterized metabolites by automated procedures. Even if the approaches cannot lead to identification of all unknown metabolites, it provides annotations (predicted pathways, network relations, potential reactions) helping to set compounds into their biochemical context.

I used the contents Chapter 3 including materials and methods, results, and conclusions to prepare a manuscript that I will submit to the Journal of Lipid Research concurrently; Quell et al. [149]. Large parts of Chapter 4 and respective materials and methods are based on and significantly extend (e.g. through replication, automated candidate selection, and database organization concepts) my paper published in the Journal of Chromatography B; Quell et al. 2017 [150]. The references are specified explicitly in the beginning of respective chapters. Both main Chapters 3 and 4 consist of independent results and discussions, but share common materials and methods in Chapter 2. In addition, a common discussion of both main chapters is provided in the summary (Chapter 5). In this thesis external references within the text refer to one or few sentences and citations at the end of a paragraph refer to the contents of the respective paragraph.

CHAPTER 2

Materials and methods

This chapter introduces cohorts, measuring devices and procedures, and describes workflows for (*i*) characterization of phosphatidylcholine (PC) compositions that have been measured as lipid sums and for (*ii*) annotation of unknown metabolites measured by non-targeted MS devices. In general, analyses were performed, and related plots were generated by R project version 3.4.0 [151].

2.1 Characterization of PC compositions

PCs measured on the lipid species level consist of several PCs of the fatty acid level. We used samples from a cohort study that have been measured on two platforms covering both precision levels. First, we systematically collected all possible constituents per PC-sum and in a second step, we estimated quantitative compositions. For replication, we tested the variation of resulting compositions within an independent cohort and across samples. With most contents of this section I also prepared a manuscript for concurrent submission; Quell et al. [149].

2.1.1 Human plasma samples of a controlled cohort

Plasma samples were collected in the HuMet study [25]. The cohort consists of 15 healthy male subjects that have been selected based on following criteria: young age (22 − 33y), no medication, and no abnormalities in standard clinical parameters. Participants were submitted to a series of different challenges over four days: extended (36h) fasting (FASTING) ingesting only 2.7 liters of mineral water, standard liquid mixed meal (standard liquid diet (SLD)) receiving a fiber-free formula drink supplying one third of their individual recommended daily energy, oral glucose tolerance test (OGTT) ingesting a commercial solution of 75 g glucose, physical exercise (physical activity test (PAT)) exercising 30 min on a bicycle ergometer at their personal anaerobic power level, oral lipid tolerance test (OLTT) ingesting SLD plus a mixture of 35 g lipids per square meter body surface area and stress test (STRESS) immersing one hand in ice water for a maximum of three minutes. Blood plasma samples were drawn at 56 different time points before, during and after challenges.

Plasma metabolite levels from both metabolomics platforms were available for the samples of four subjects (subject 5 – subject 8) except one time point for subject 7 leading to a total number of 223 samples.

The Qatar Metabolomics Study on Diabetes (QMDiab), in which we sought replication of our results, consists of 196 diabetic and 202 non-diabetic subjects aged between 17 and 82 years. Metabolite levels of 305 non-fasting blood plasma samples (152 male, 153 female) measured on the same two metabolomics platforms were available [152].

2.1.2 PC quantification on the lipid species level (Absolute*IDQ*$^{\text{TM}}$)

For quantifying lipid sums, HuMet and QMDiab plasma samples were analyzed using the Absolute*IDQ*$^{\text{TM}}$ p150 kit (Biocrates$^{\text{TM}}$ Life Sciences AG, Innsbruck, Austria). Besides PCs (including lysoPCs) and SMs, carnitines, amino acids, and hexoses are quantified by this targeted metabolomics approach. The corresponding analytical procedures have been described in detail before [25, 78]. In brief, 10 µl of human plasma were pipetted onto filter inserts (containing internal standards) in a 96 well plate. The filters were tried under a nitrogen stream, amino acids were derivatized by addition of a phenylisothiocyanate reagent (5%), and samples were dried again. After extraction with 5 mM ammonium acetate in methanol, the solution was centrifuged through a filter membrane and diluted with running solvent. Metabolites were detected by direct infusion to a 4000 QTRAP system (Sciex$^{\text{TM}}$, Darmstadt, Germany) equipped with a Shimadzu Prominence LC20AD pump and SIL-20AC auto sampler. 76 PC, 14 lysoPC, and 15 SM species were measured in positive MRM scan mode selective for the common fragment ion of the PC head group (m/z = 184). The isobaric metabolites (within the mass resolution of the MS device) PC $x : y$ and PC O-$x+1 : y$ cannot be distinguished by this technique. Since odd chain lengths are considered rare for free fatty acids, measured PCs are principally labeled under the assumption of an even number of carbon atoms in the fatty acid chains, i.e., PC aa C$x : y$ (for PC $x : y$) is chosen as label in case of even x and PC ae C$x + 1 : y$ (corresponding to PC O-$x + 1 : y$) in case of odd x. In total, four internal standards were used for quantification of the phospholipid species (each one for lysoPC, SM, short chain PC, and long chain PC). Concentrations are calculated by the Absolute*IDQ*$^{\text{TM}}$ kit software and reported in µmol/L. During QC, metabolites with a coefficient of variation (CV) above 25% and metabolites that showed a significant correlation to the run day and had a CV above 20% were excluded, preserving 68 and 72 PC, 8 and 10 lysoPC, and 13 and 14 SM species of HuMet and QMDiab for further analyses.

2.1.3 PC quantification on the fatty acid level (Lipidyzer$^{\text{TM}}$)

For quantifying lipids on the fatty acid level, plasma samples were analyzed on the Lipidyzer$^{\text{TM}}$ platform of AB Sciex$^{\text{TM}}$ Pte. Ltd., Framingham, USA, and Metabolon$^{\text{TM}}$ Inc., Durham, NC, USA. The method allows quantification of over 1 100 lipid species from 14 lipid subclasses, including 18 lysoPCs, 109 PCs and 12 SMs [153–156]. Lipids were extracted from samples using dichloromethane and methanol in a modified Bligh-Dyer extraction in the presence of internal standards with the lower, organic phase being used for analysis [157]. The extracts were concentrated under nitrogen and reconstituted in 0.25 mL of dichloromethane:methanol (50:50) containing 10 mM ammonium acetate. The extracts were placed in vials for infusion-MS analyses, performed on a Sciex$^{\text{TM}}$ 5500 QTRAP equipped with the SelexION$^{\text{TM}}$ differential ion mobility spectrometry (DMS) technology [158, 159]. This additional device added to the ESI source (applied in positive and negative mode) allows scanning or filtering molecules for their ion mobility in alternating high and low electric fields before entering the mass spectrometer. With this technology a separation of lipid classes (in PCs, lysoPCs, SMs, etc.), independent of their mass-to-charge ratio (m/z) is achieved. DMS-MS conditions were optimized for lipid classes. PCs are detected in negative MRM mode, with characteristic mass fragments for the fatty acid side chains, thus allowing a resolution of the fatty acid composition. Individual lipid species were quantified based on the ratio of signal intensity for target compounds to the signal intensity for an assigned internal standard of known concentration. For quantification of PCs, 10 labeled compounds with a C16:0 fatty acid side chain in the *sn*-1 position and side chains in the range of C16:1 − C22:6 in the *sn*-2 position were used as internal standards [154, 160, 161].

2.1.4 Qualitative description of PC sums

For assigning PCs measured by Lipidyzer$^{\text{TM}}$ to Absolute*IDQ*$^{\text{TM}}$ PC-sums PC $x : y$ (annotated with PC aa C$x : y$ and PC ae C$x : y$ in the kit), we first collected all theoretically possible lipid isobars of PC $x : y$ of the fatty acid level. To this end, we systematically distributed the total numbers of carbon atoms x and double bonds y to the two fatty acid chains, using the short hand notation PC $x_1 : y_1\_x_2 : y_2$ with $x = x_1 + x_2$ and $y = y_1 + y_2$ to indicate that we do not distinguish between the *sn*-1 and *sn*-2 positions of the side chains [162]. Additionally, the isobaric compounds PC $x+1 : y+7$, PC O-$x+1 : y$ and PC O-$x+2 : y+7$ were considered for PC $x : y$. Notably, we also included PCs containing odd chain fatty acids or fatty acids with very high desaturation to generate a comprehensive list of isobaric PCs. The software coming with the Absolute*IDQ*$^{\text{TM}}$ kit, corrects concentrations of PC-sums for the concentration of the (within the mass resolution) isobaric SMs [^{13}C$_1$]SM x+4 : y, if the corresponding SM has been quantified. If the corresponding SM was not quantified by the kit, i.e., no isotope correction has been performed, we included [^{13}C$_1$]SM $x+4 : y$ in the list of possible lipid species measured under the PC $x : y$ sum.

2.1.5 Quantitative estimation of PC compositions

Both platforms report concentrations of lipid measures in µmol/L. Therefore, to determine the fraction of a PC of the fatty acid level (measured on the Lipidyzer$^{\text{TM}}$ platform) within a PC-sum (obtained by the Absolute*IDQ*$^{\text{TM}}$ kit), factors f, we divided the concentration of the PC j (Lipidyzer$^{\text{TM}}$) by the sum measure that includes this PC, PC i (Absolute*IDQ*$^{\text{TM}}$), according to our qualitative assignment (Equation 2.1).

$$f_{ij} = \frac{\sum_{s \in \text{Subjects}} \sum_{t \in \text{Timepoints}} q_{ijst}}{\mid \text{Subjects} \mid + \mid \text{Timepoints} \mid} \tag{a}$$

$$\text{with} \quad q_{ijst} = \frac{conc.(PC_{j:\text{fatty acid level}_{st}})}{conc.(PC_{i:\text{lipid species level}_{st}})} \tag{b}$$

Equation 2.1: Factors *f* describing quantitative compositions of PCs
Factors f_{ij} describe the mean value of quotients q_{ijst} of all subjects and time points. Underlying quotients q_{ijst} estimate the quotient of a PC j measured on the fatty acid level and the respective PC i measured on the lipid species level for a specific subject s and time point t.

The variation of factors f_{ij} was specified as 5% and 95% confidence interval, in which the 5% interval contains the lowest 5% and the 95% interval the largest 5% of the ratio values q_{ijst}. Additionally, we determined fractions of isobaric PCs of the fatty acid level with no regard to the respective PC of the lipid species level or non-measured constituents (R): e.g. PC 16:0_20:2 / (PC 16:0_20:2 + PC 18:0_18:2 + PC 18:1_18:1).

2.1.6 Replication of estimated quantitative PC compositions

Factors f_{ij} (estimating the ratio of PCs measured on the fatty acid level (j) and the respective PC measured on the lipid species level (i): $PC_{i:\text{lipid species level}} \cdot f_{ij} = PC_{j:\text{fatty acid level}}$) were replicated with metabolite levels of the subset of 31 male controls aged between 17 and 40 years of the independent QMDiab cohort (to receive subjects with comparable properties to participants of the HuMet cohort). After running the cross-platform comparison of the selected QMDiab samples (as described above), a Kruskal-Wallis test (function '`kruskal.test`' of R package '*stats*' version 3.4.0 [151]) tested, if the distributions of quotients q (factors f correspond to the mean value of quotients q across all subjects and time points, cf. Equation 2.1) of HuMet- or QMDiab-sujects are significantly different (significance level $\alpha = 0.01$). If distributions of quotients q of the main compound (i.e., the constituent with the largest f) of a composition are not significantly different, the composition was labeled as replicated.

2.1.7 Imputation of measured PC levels

As complementary analysis, we used factors f determined in HuMet and PCs (lipid species level) of male controls aged ≤ 40y of QMDiab to impute PCs of the fatty acid level: $PC_{i:\text{lipid species level (QMDiab)}} \cdot f_{ij(HuMet)} = PC_{j:\text{fatty acid level (Imputed)}}$. A Kruskal-Wallis test (function '`kruskal.test`' of R package '*stats*' version 3.4.0 [151]) showed whether distributions of observed and imputed concentrations of PCs (fatty acid level) in QMDiab are significantly different ($\alpha = 0.01$). Imputation of PC compositions was labeled as successful, if the distributions of concentrations of their main constituents were not significantly different.

2.1.8 Testing variation of compositions across subjects and challenges

The variation of estimated factors f_{ij} was tested for each PC j measured on the fatty acid level and the respective PC-sum i (lipid species level) across subjects and during challenges.

For each subject s in HuMet, a Shapiro-Wilk test (function '`shapiro.test`' of R package '*stats*' version 3.4.0 [151]) was applied to check, if q_{ijst} is normally distributed within 56 time points t. For normally distributed ratios, an analysis of variance (ANOVA) (function '`aov`' of R package '*stats*' version 3.4.0 [151]), otherwise a Kruskal-Wallis test (function '`kruskal.test`' of R package '*stats*' version 3.4.0 [151]) was performed to test if quotients q originate from the same distribution. If the null hypothesis with an α-level of 0.01 could not be rejected, distributions were considered as stable over samples.

In addition, cross-platform comparisons were run using three subsets of the QMDiab cohort: male controls, female controls and cases. A Kruskal-Wallis test was used to check, if factors f (distributions of quotients q) between those three groups were significantly different.

To investigate the stability of factors f during challenges, we used metabolite levels of the three HuMet challenges with relation to lipid metabolism, namely FASTING, PAT, and OLTT. For each challenge, we considered 'baseline' samples denoting samples drawn before the challenges and 'intervention' samples at the time point with largest changes of the total metabolome compared to baseline: FASTING: baseline at 8 am after 12h overnight fasting; intervention at 36h fasting; PAT: baseline at 4 pm (4h after last meal); intervention after 30 min exercise; OLTT: baseline at 8 am after 12h overnight fasting; intervention at noon.

Central trends of quotients q for baseline and intervention time points were substantiated by a Wilcoxon signed-rank test (function '`wilcox.test`' with options 'two sided' and 'paired' of R package '*stats*' version 3.4.0 [151]), since both time points consist of a maximum of four samples. Even in case of large differences between distributions of quotients q of baseline and intervention, the null hypothesis was not rejected at an α-level of 0.05.

2.1.9 Estimation of unmapped constituents of PC sums

To estimate to what extent the PC-sum measure is explained by all measured PC species of the fatty acid level, we applied linear models and Bland-Altman analyses.

The formulas for linear models were constructed by

$$PC_{\text{lipid species level}} \sim b \cdot \sum PC_{\text{fatty acid level}}$$

and evaluated by the ‘`lm`’ function of the R package ‘*stats*’, version 3.4.0 [151]. Thereby, PCs of the fatty acid level were weighted equally within one formula. In the Bland-Altman plots, the difference of PC-sums (lipid species level) and cumulated respective PCs of the fatty acid level

$$conc.\left(PC_{\text{lipid species level}}\right) - \sum conc.\left(PC_{\text{fatty acid level}}\right)$$

were opposed to the mean of concentrations

$$\left[conc.\left(PC_{\text{lipid species level}}\right) + \sum conc.\left(PC_{\text{fatty acid level}}\right)\right]/2$$

of both platforms.

2.2 Annotation of metabolites measured by non-targeted metabolomics

Non-targeted MS techniques incorporate measurements of metabolites, whose chemical structures are unknown. Besides identification of the molecular structure of an unknown metabolite, annotation of these metabolites includes for instance classification into biochemical pathways and determination of substrates or products of related enzymatic reations to elucidate their biochemical function. The modules of our automated approach characterize these unknown metabolites by estimation of GGMs, integration of generated and public data, pathway and reaction prediction, and automated selection of specific candidate molecules. Figure 1.3 and Figure B.1 show an overview of the complete workflow. We demonstrated the applicability of our method on metabolomics data that was produced by a non-targeted LC-MS analysis in the course of the SUMMIT project and on further data of cohort studies including KORA F4, TwinsUK, or GCKD. Implementations of all modules in R are provided in Supplemental File B.1 along with the data on which the here presented analyses are based. Candidate molecules that our method predicted for selected unknown metabolites were confirmed (or excluded) by an experimental validation. I previously published (parts of) materials and methods described in Sections 2.2.1 to 2.2.6, 2.2.8, and 2.2.10 in Quell et al. 2017 [150].

2.2.1 Study cohorts

Serum samples of n = 2 279 patients with type 2 diabetes from seven population studies, FINRISK1997 (n = 242), FINRISK2002 (n = 92), FINRISK2007 (n = 28) [19], Go-DARTS (n = 1 200) [20], IMPROVE (n = 44) [21], 60 years-olds (n = 20) [22], and SDR (n = 653) [23], all participating in the SUMMIT project, were analyzed using the commonly used (established) non-targeted metabolomics platform of Metabolon$^{\text{TM}}$ Inc. (Durham, USA). 1 147 of the type 2 diabetes patients were also diagnosed with cardiovascular disease, while 1 132 did not suffer from cardiovascular disease. Besides the type 2 diabetes and cardiovascular disease state, patients provided further clinical information such as age, sex, duration of type 2 diabetes, height, BMI, triglyceride, HDL, LDL, DBP, SBP, smoking status, hemoglobin A1c, baseline estimated glomerular filtration rate, insulin status, and medication information including ACE inhibitors, angiotensin receptor blockers, calcium channel blockers, diuretics, lipid rx, blood pressure lowering drugs, beta blockers, alpha blockers, and aspirin.

As second data set, we used metabolite levels measured on the new high-resolution non-targeted LC-MS platform of Metabolon$^{\text{TM}}$ in 1 209 samples of the GCKD study to demonstrate our module for automated selection of candidate molecules. In total, the GCKD cohort consists of about 5 000 patients (aged 17 – 74y) with chronic kidney disease of various aetiologies. Patients are under nephrological care and their pathogeny will be followed during a 10-year period, in which matrix samples (blood serum, plasma, urine) are collected at regular intervals. Besides matrix samples, patients provide information concerning their demography, anthropometric data, renal and cardiovascular history, renal biopsy history, comorbidities, medication, life style, family history, heart failure, angina pectoris/dyspnoea, and intermittent claudication. [18]

2.2.2 Metabolomics measurements

The non-targeted metabolomics platform comprises LC-MS (in positive and negative mode) as well as MS coupled to GC and has been described in detail previously [163, 164]. Briefly, samples were thawed on ice and extracted with methanol containing internal standards to control extraction efficiency. Extracts were split into aliquots for positive and negative LC-MS and GC-MS mode and dried under nitrogen. LC-MS analyses were performed on an LTQ XL mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA, USA) coupled to a Waters Acquity UPLC system (Waters Corporation, Milford, MA, USA). For LC-MS positive (negative) ion analysis 0.1% formic acid (6.5 mM ammonium bicarbonate [pH 8.0]) in water was used as solvent A and 0.1% formic acid in methanol (6.5 mM ammonium bicarbonate in 95% methanol) as solvent B. After sample reconstitution with solvent A and injection, the column (2.1 mm × 100 mm Waters BEH C18, 1.7 µm particle-size) was developed with a gradient of 99.5% solvent A to 98% solvent B. The flow rate was set to 350 µL/min for a run time of 11 minutes each. The eluent was directly connected to the ESI source of the mass spectrometer. Full MS scans were recorded from 80 to 1 000 m/z, alternating with data dependent MS/MS fragmentation scans with dynamic exclusion. GC-MS analyses were performed on a Finnigan Trace DSQ single quadrupole mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA, USA) containing a GC column (20 m × 0.18 mm, 1.8 µm film phase consisting of 5% phenyldimethyl silicone). The GC was performed during a temperature gradient from 60 to 340°C with helium as carrier gas. MS scans with electron impact ionization (70 eV) and a 50 to 750 m/z scan range were used. The metabolite identification has been semi-automated performed by Metabolon^{TM} Inc. using a reference spectra library. Further details for the LC-MS part are also given below the description of the experimental validation of the selected candidates. [163, 164]

The non-targeted high-resolution platform of Metabolon^{TM} performs LC-MS measurements (in positive and negative mode) with a mass resolution of 5 ppm. In addition to the procedure described above, chemical separation was performed on a hydrophilic interaction liquid chromatography (HILIC) system 2.1 mm × 150 mm Waters BEH Amide 1.7 µm (at a temperature of 40°C). The column was developed with a gradient from 95% solvent A (10 mM ammonium formate in 15% water, 5% methanol, and 80% acetonitrile [pH 10.16]) to 95% solvent B (10 mM ammonium formate in 50% water, 50% acetonitrile [pH 10.60]) within 5.5 min. and back to 95% solvent A until min. 11. The flow rate was set to 500 µL/min. Up to nine full MS scans per second (repeated sequences of MS and MS/MS scans) were recorded from 80 to 1 000 m/z on a Q-Exactive (orbitrap) mass spectrometer (Thermo Fisher Scientific Inc.). [165, 166].

In total, the levels of 758 metabolites were determined for the 2 279 subjects in our cohorts. For 319 of the metabolites the chemical identity was not known at time of analysis. The 439 known metabolites are assigned to a simplified two-level metabolite ontology, consisting of 8 super pathways and 102 more precise sub pathways, which is similar to the ontology used by KEGG [38]. In 1 209 samples of a second cohort, 1 456 metabolites (796 known, 660 unknown) were quantified by the high-resolution non-targeted LC-MS platform of Metabolon$^{\mathrm{TM}}$.

2.2.3 Public data sources and data types

For integrating metabolite-metabolite and metabolite-gene links based on known biochemical reactions into our model, we used Recon (Virtual Metabolic Human Database) [110], a community driven reconstruction of the human metabolism incorporating reactions between metabolites and functional gene annotations, as a representative among available metabolic databases including KEGG [38] or HumanCyc [115]. Furthermore, we used a published GGM based on the population cohort KORA F4 (n = 1 768) [147], and metabolite-gene associations of a published mGWAS based on KORA F4 and TwinsUK (n = 6 056) [167].

Our module for automated candidate selection requires access to chemical structures of public data bases. We used 1 995 Recon-structures [110] and 112 770 HMDB-structures [71], provided as chemical fingerprints or 2D-substructures (cf. Chapter 2.2.7).

2.2.4 Data processing and integration

Preprocessing

Metabolite concentrations were normalized by the median per metabolite and run day. Afterwards, metabolite concentrations were Gaussianized, meaning that values per metabolite were sorted and transferred to values of a normal distribution [168].

GGM generation

Based on the metabolomics data of the 2 279 subjects of our SUMMIT-cohorts, we created a GGM as backbone of our network model, since they are known to reconstruct biochemical pathways from measured metabolite levels [118]. First, we excluded metabolites with more than 20% and samples with more than 10% missing values leaving 625 metabolites for GGM generation. To generate a complete data matrix as required for the following analyses, we utilized the R package '*mice*' [169] with standard parameters to impute the remaining missing values. '*mice*' implements algorithms for multivariate imputation by chained equations. The standard method 'pmm' (predictive mean matching) in the mice-package estimates missing values of a variable x by applying regression models incorporating all other variables in the input matrix. A missing value in x is finally imputed as the value belonging to one of the 5 cases with observed values in x, for which the value that is predicted based on the regression is closest to the predicted value of the case with missing data on x. To calculate partial correlations between metabolites in the complete data matrix, which form the basis of GGMs, we applied the R package '*GeneNet*' [170] with the function '`ggm.estimate.pcor`' and the method 'dynamic'. The function '`network.test.edges`' extracted 862 significant GGM edges according to the Bonferroni corrected threshold of 0.01. To avoid biases in the network model related to covariates that are known or suspected to influence metabolite levels, in each calculation sex, age, study, and the clinical phenotypes mentioned above were considered as covariates by including them into the input data matrix for '`ggm.estimate.pcor`'.

Data integration

We merged the edges of the newly generated GGM with 398 significant partial correlations of the published GGM based on the KORA F4 cohort [147] into one model to end up with 1 040 connections between 637 known and unknown metabolites. Then we added 134 metabolite-associated genes of a published mGWAS [167]. Finally, we attached knowledge-based biochemical information extracted from Recon [110] to the network model. To this end, we first added metabolites from Recon if they were functionally related to at least one of the 134 genes through a reaction listed in Recon. 343 Recon metabolites, of which 37 were mapped to measured metabolites and thus were already part of the GGM- and mGWAS-based network, showed functional links to 57 genes, which were added to the network. Secondly, Recon reactions between metabolites annotated as '*baseReactants*' and '*baseProducts*' in the Recon data file were attached to the network if at least one of these metabolites could be mapped onto a measured known metabolite. Following this procedure, we found reactions for 83 measured metabolites and included them into the network. Thereby, 174 metabolites were added. In the final step, we complemented the network by incorporating edges between all metabolites in the network that were connected by a Recon reaction. In total, the resulting (final) network includes, 1 152 Recon reactions connecting 591 metabolites. Please note that by integrating only main metabolites, which are annotated as '*baseReactants*' and '*baseProducts*' in Recon, we avoid connecting metabolites via so-called side metabolites (e.g. cofactors, water) which would lead to biochemically incorrect edges in the network. While Recon provides annotation concerning main and side metabolites making the role of metabolites in a reaction directly accessible, more sophisticated methods (e.g. using chemical similarity between metabolites) are needed if knowledge on biochemical reactions is extracted from other resources that do not include such annotations [171–173]. Please also note that, for our purposes, we ignored the compartment annotations provided with metabolite species in Recon reactions, i.e., each Recon metabolite mapped or added to the GGM- and mGWAS-based network is represented by a single node and two metabolites are connected if they are linked through a Recon reaction irrespective of the compartment in which the reaction takes place. Reactions that are classified as '*Transport*' or '*Exchange*' in Recon are omitted when integrating Recon reactions into the network. The data integration process is schematically visualized in Figure B.2. The 'organic' layout of the yEd Graph Editor (yWorks GmbH, Tübingen) was performed to distribute nodes of the visual representations of network models (*graphml*).

2.2.5 Automated prediction of super and sub pathways

Each known metabolite can be annotated using one of several existing metabolite ontologies (pathway schemes). Estimating the assignment within the ontology for unknown metabolites helps to shrink the list of possible candidate molecules using the biochemical context of unknown metabolites. Here, we are using the annotation that was provided with the metabolomics data, which assigns each metabolite to one of 8 non-overlapping super pathways and a more precise sub pathway. Any other classification scheme could be applied analogously within our method. While, in general, more fine-grained, and thus more specific pathway definitions can be expected to allow more precise predictions, they will, at the same time, produce more ambiguous pathway assignments for unknown metabolites, in particular in case of overlapping pathways where a metabolite can be annotated with various pathways.

The idea of our approach is to capture the neighborhood of each known metabolite and to count the frequencies of their pathways. These frequencies can then be used to estimate the most probable pathway for each unknown metabolite considering the pathways of the known metabolites in its neighborhood (Figure 4.2(b)). To define the neighborhood for metabolites, we consider metabolites as neighbors, if they are connected by a GGM edge, share a common mGWAS gene, if there is a gene associated to the unknown metabolite and this gene is functionally related to a known metabolite, or if the unknown metabolite is connected by a GGM edge to a known metabolite, which is connected through a reaction to a database metabolite (Figure 4.2(a)).

Our approach is divided into a training phase based on known metabolites and a prediction phase, in which the super pathway p_i is predicted for each unknown metabolite i. During the training phase we first determined the a priori probabilities $P_B(p)$ of each super pathway $p \in \{$Amino acid, Carbohydrate, Cofactor and vitamins, Energy, Lipid, Nucleotide, Peptide, Xenobiotics$\}$ (Equation 2.2).

$$P_B(p) = \frac{\sum \#\text{neighbors of metabolites with super pathway } p}{2 \cdot \#\text{neighboring metabolite pairs}}$$

Equation 2.2: A priori probability of pathway *p*
The frequency of pathway p is set into relation to the connectivity (summed node degrees) of metabolites annotated with p in the network model.

For each super pathway p, we calculated the conditional probability $P_N(p_i \mid p_j)$ based on all known metabolites i with super pathway p_i given metabolites j with super pathway p_j are neighbors of i (Equation 2.3).

$$P_N(p_i \mid p_j) = \frac{P(p_i \cap p_j)}{P_B(p_j)} = \frac{P(p_i \text{ and } p_j \text{ are neighbors})}{\text{background probability of } p_j}$$

Equation 2.3: Probability of neighboring metabolites having pathways p_i and p_j
The ratio estimates the conditional probability that metabolites with pathway p_i have neighbors p_j.

During the prediction phase we resolve the conditional probability $P_N(p_i \mid p_1 \cap \ldots \cap p_n)$ for each super pathway p_i of all unknown metabolites i given the pathways $p_1, \ldots, p_n$ of n neighboring known metabolites (Equation 2.4). The first transformation follows the Bayes theorem. For the second transformation we assumed independence of the neighbors $1, \ldots, n$. As a consequence of the approximation, a very small value of one conditional probability results in a very small overall probability for the specific pathway.

$$
\begin{aligned}
P_N(p_i \mid p_1 \cap \ldots \cap p_n) \Leftrightarrow & \frac{P_N(p_1 \cap \ldots \cap p_n \mid p_i) \cdot P_B(p_i)}{P_B(p_1 \cap \ldots \cap p_n)} \\
\sim\propto & \frac{P_N(p_1 \mid p_i) \cdot \ldots \cdot P_N(p_n \mid p_i) \cdot P_B(p_i)}{P_B(p_1) \cdot \ldots \cdot P_B(p_n)}
\end{aligned}
$$

Equation 2.4: Prediction of pathway p_i for metabolite surrounded by p_1 to p_n
The formula estimates the conditional probability that a metabolite has pathway p_i given that its neighbors have pathways p_i to p_n. The probability is estimated for each pathway, keeping the pathway with the largest probability.

For each unknown metabolite i, the predicted super pathway p_i with the highest probability $max(P_N(p_i \mid p_1 \cap \ldots \cap p_n))$ is accepted if its probability is at least z times higher than the super pathway with the second highest probability. We defined five classes of confidence and estimated respective values of z empirically based on multiple 10-fold cross-validations with known metabolites: (a) very high confidence (correct predictions $\geq 97.5\% \Rightarrow z \geq 207.0$), (b) high confidence (correct predictions $\geq 95\% \Rightarrow z \geq 78.0$), (c) medium confidence (correct predictions $\geq 90\% \Rightarrow z \geq 7.1$), (d) low confidence (correct predictions $\geq 85\% \Rightarrow z \geq 2.7$) and (e) very low confidence (correct predictions $< 85\% \Rightarrow z < 2.7$). For metabolites, that are neighbors of further unknown metabolites, but not of known metabolites, we used the super pathway with the highest a priori probability and assigned the confidence class according to the criteria above.

For each unknown metabolite with a predicted super pathway, we selected the more specific sub pathway that is most common among neighboring known metabolites. If an equal number of neighbors own different sub pathways, we stored each option.

We evaluated our approach with a series of 100 10-fold cross-validations using known metabolites with annotated super and sub pathways.

2.2.6 Automated prediction of reactions connecting metabolites

Knowledge about the reaction connecting a known to an unknown metabolite leads to the possibility of virtually applying this reaction to the known metabolite to select candidate molecules. Here, we focused on the 21 frequently occurring reaction types that are shown in Table 4.3. We assumed the presence of a reaction between two metabolites if they were connected directly by a GGM edge or indirectly via a gene based on a mGWAS or via a known reaction according to Recon (Figure 4.2(c)). This assumption is based on the observation that pairs of metabolites that are connected by a GGM edge particularly tend to be reactants of a direct reaction [118]. Nevertheless, direct edges in the GGM might represent also multi-step-reactions between two metabolites in cases where the intermediate metabolites are not quantified. Following a simplified approach, we then assigned a specific reaction type to a connected pair of metabolites, if the two metabolites showed an m/z difference indicating a difference in molecule mass that is typical for the respective reaction type with $\Delta m_{pair} = \Delta m_{expected} \pm e$ [148]. Here, we set $e = 0.3$ to compensate the unit mass resolution of the MS platform, on which the presented data was collected. We adapted e to 0.01 for our second data set measured on a high-resolution MS platform to yield more specific reaction types.

To increase the specificity and to shrink the list of assigned reactions, we additionally filtered the list by differences in RI as a second criterion for the most frequent reaction type, the dehydrogenation reaction. The distribution of differences in RI between all GGM pairs of known metabolites with correctly predicted dehydrogenation reaction (26 true positives) is compared to the respective distribution for wrongly predicted dehydrogenation reaction (8 false positives). Since both distributions do not overlap (Figure B.3), we used the mean difference in RI of the correct predictions plus/ minus their variances ($\Delta\text{RI} \leq 355.9$) as threshold.

Predicted reactions can be applied to known metabolites to manually select specific candidate molecules for the connected unknown metabolites that can be verified experimentally.

2.2.7 Comparative annotation of incorporating different platforms

A second data set (GCKD) allowed (*i*) replication of automated annotations and (*ii*) assessment of benefits for applying our workflow with merged data sets. Even though samples of SUMMIT were measured by the established Metabolon^{TM} platform, and samples of GCKD were measured on the new high-resolution platform of Metabolon^{TM}, a subset of 227 metabolites (168 known, 59 unknown) was mapped and therefore available in both sets.

Merging multiple automated annotations

For comparing automated annotations, we applied our data integration module as described in Chapter 2.2.4. To this end, we filtered levels of 1 382 metabolites and 1 209 samples according to a maximum of 20% missing values on the metabolite dimension, and 10% missing values on the metabolite dimension. Levels of remaining 986 metabolites (496 known, 490 unknown) were imputed by 'mice' (R package 'mice' [169]) and forwarded to estimation of 1 083 significant pairwise partial correlations with age, sex, and BMI as covariates (R package '*GeneNet*' [170]). The resulting GGM and 714 published genome wide significant metabolite-gene associations ($\text{pv} \leq 1 \cdot 10^{-6}$; mGWAS of the GCKD study [174]) were integrated with reactions and functional metabolite-gene relations of Recon [110] (cf. Chapter 2.2.4). The modules for automated prediction of pathways and reactions were applied on the generated network model as described in Chapters 2.2.5, and 2.2.6.

For replication, we considered a subset of 59 unknown metabolites that have been measured in both sets (samples of SUMMIT measured on the established Metabolon$^{\text{TM}}$ platform and samples of GCKD measured on the new high-resolution platform of Metabolon$^{\text{TM}}$), and counted equally predicted super and sub pathways. In case of sub pathways, comparable predictions (e.g. 'short chain fatty acid' and 'medium chain fatty acid'; 'androgenic steroids' and 'sterol, steroid'; or 'fatty acid metabolism' and 'long chain fatty acid') were labeled as similar. Finally, we recognized different a priori probabilities of super pathways among both sets of known metabolites.

Merging data sets for a common automated annotation

In a second step, we searched for benefits resulting from merging both data sets, before applying the automated annotation modules. To this end, we mapped 227 metabolites (168 known, 59 unknown) that were measured in samples of SUMMIT (established Metabolon$^{\text{TM}}$ platform) and GCKD (new high-resolution platform of Metabolon$^{\text{TM}}$) based on their Metabolon$^{\text{TM}}$-ID, Name, and HMDB-ID. Significant partial correlations (GGM) were generated within both sets separately (described previous sections), to prevent biases, and resulting 2 043 GGM edges were subsequently integrated by the automated workflow (cf. Chapter 2.2.4). Furthermore, 978 metabolite-gene associations of published mGWAS [167, 174] as well as reactions and functional relations of Recon were added to the network model. Finally, we applied our modules for automated prediction of pathways and reactions to the combined network model 'SUMMIT/GCKD' (cf. Chapters 2.2.5 and 2.2.6).

During this analysis we first compared super and sub pathways of 427 unknown metabolites that were predicted in at least two of the three sets to test if automated annotations are influenced by changed network models: 'SUMMIT/GCKD', SUMMIT, and GCKD (procedure described in the previous section). Special interest was on a subset of 59 unknown metabolites that were measured in both sets. The second focus of this analysis was on 8 unknown metabolites that were automatically annotated based on the merged model but could not be annotated based on both separate models.

2.2.8 Computer aided manual candidate selection

Automated annotations facilitate the manually selection of specific candidate molecules for unknown metabolites. In case of metabolites measured by the established Metabolon^{TM} platform, we used our predicted pathways for a first classification of unknown metabolites. Furthermore, we manually applied predicted reactions (e.g. especially dehydrogenation reactions that passed the ΔRI criterion) to neighboring known metabolites by changing their structures (e.g. introduction or removal of a double bound) to receive molecules matching the measured mass of respective unknown metabolites. Public data bases (e.g. Recon [110], HMDB [71], ChemSpider [73], LipidMaps [175], PubChem [72], or KEGG [38]) were searched to check, whether selected candidates have already been annotated in biochemical contexts.

Especially for metabolites measured on the new high-resolution platform of Metabolon^{TM}, we additionally searched HMDB [71] for structures with monoisotopic masses corresponding to the measured mass (m/z ± 0.01, or ± 0.3 for the established Metabolon^{TM} platform) and examined the plausibility of the structure in the biochemical context of the network model.

2.2.9 Automated candidate selection procedure

For unknown metabolites that were measured on high mass resolution, we established an automated candidate selection procedure. This procedure makes use of an increased structural similarity of neighboring (based on the network model) metabolites compared to random pairs. Potential candidate molecules can be selected within metabolite-pools of databases such as Recon [110] or HMDB [71].

Integration of chemical structures

Besides the network model described in Chapters 2.2.4 to 2.2.6 this procedure requires a candidate pool containing metabolites of databases including names, neutral masses and structural information. For 1 995 of 2 625 Recon-compounds, the extensible markup language (XML)-based structure data files (SDFs) including chemical annotations of molecules were obtained from the PubChem download service [72]. The metabolite names, their database ID and neutral monoisotopic masses were read or calculated from the SDFs (fields '*PUBCHEM_IUPAC_NAME*', '*PUBCHEM_COMPOUND_CID*', '*PUBCHEM_MONOISOTOPIC_WEIGHT*' and '*PUBCHEM_TOTAL_CHARGE*') and stored as separate searchable table by our model to build a candidate pool [176].

For selection of candidates for unknown metabolites, 112 770 molecules of HMDB version 4.0 were added to the candidate pool [71]. Information was read from the fields '*JCHEM_TRADITIONAL_IUPAC*', '*HMDB_ID*' and '*EXACT_MASS*' for each compound.

Chemical fingerprints of PubChem compounds

Structural descriptors such as chemical fingerprints and maximum common substructures (MCSs) were used to classify and systematically compare molecules on the basis of common elements. Chemical fingerprints represent typical properties and substructures of metabolites within a standardized format. [177]

The PubChem fingerprint consists of a binary 881 bit tuple. Each bit indicates whether a specific substructure is part of the target molecule (*on-bit*), or not (*off-bit*). Some bits contain quantitative information about substructures: While bits 0 – 114 refer to single atoms (e.g. '*>= 4 C*'), bits 115 – 262 incorporate the size, number, and saturation of carbon-, nitrogen-, or heteroatom-containing ring structures (e.g. '*>= 2 saturated or aromatic carbon-only ring size 3*'). Bits 263 – 326 test the existence of simple pairs of bonded atoms with no regard to their count (e.g. '*C–O*'). Bits 327 – 415 and bits 416 – 459 commit nearest neighboring groups of up to five atoms without (~) and with regard to their bond order (single (–), double (=), or triple (#) bond order) (e.g. '*C(~O)(~O)*' or '*C(–O)(=O)*'). Finally, bits 460 – 712 and bits 713 – 880 provide information about simple and complex SMILES arbitrary target specification (SMARTS) patterns, respectively, including specific bond orders (e.g. '*O–C–C=N*' or '*Cc1ccc(C)cc1*'). [176]

Maximum common substructure of chemical structures

Besides structural descriptor based chemical fingerprints, MCSs enable a direct determination of the largest coherent substructure (number of equally connected atoms) that two chemical structures have in common. The function '`fmcs`' of the R-package '*fmcsR*' (version 1.0) computes the MCS of two molecules (based on SDFs) by iteratively adding common atoms and bounds of both molecules. Since finding MCSs is NP-complete, a heuristic approach based on search trees solves this task. [178]

Estimation of structural similarity

The similarity of pairwise chemical fingerprints or MCSs is quantified by Tanimoto and overlap coefficients [177–179]. In case of chemical fingerprints, the Tanimoto coefficient describes the fraction of *on-bits* that both structures have in common (Equation 2.5(a)). The number of common *on-bits* (c) is divided by the number *on-bits* of the union of both fingerprints ($a + b - c$, in which a and b are the numbers of *on-bits* of both fingerprints individually). In case of MCSs, the Tanimoto coefficient is calculated similarly: the size (atom count) of the MCS is divided by the number of atoms of the union of both structures.

As second measure, overlap coefficients quantify the fraction of the smaller structure that is contained in the larger (super-) structure (Equation 2.5(b)). Since chemical fingerprints refer to abstract elements, overlap coefficients were only calculated for MCSs. The respective atom count was used for the size of structures (a, b) and of the MCS (c).

After computation of MCSs, function '`fmcs`' of the R-package '*fmcsR*' (version 1.0) calculates their Tanimoto and overlap coefficients [178]. Tanimoto coefficients of chemical fingerprints were determined by the function '`fpSim`' of the R-package '*ChemmineR*' (version 2.30.0) [179].

$$T_{A,B} = \frac{\sum_{j=1}^{n} x_{jA} \cdot x_{jB}}{\sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 - \sum_{j=1}^{n} x_{jA} \cdot x_{jB}} \overset{\star}{\equiv} \frac{c}{a+b-c} \qquad \text{(a)}$$

$$O_{A,B} = \frac{c}{min\,(a{,}b)} \qquad \text{(b)}$$

Equation 2.5: Tanimoto and overlap coefficients
The structural similarity of molecules can be quantified by the (a) Tanimoto or (b) overlap coefficient of pairwise MCSs or chemical fingerprints respectively. $A, B :=$ the set of *on-bits* of the fingerprints or the set of atoms of both molecules, $x :=$ element of set A or B, $a := |A|$, $b := |B|$, and $c := |A \cap B|$. Equivalence$^{(\star)}$ applies for dichotomous variables. Equations adapted from: [177–179]

Selection of candidate molecules

Measuring unknown metabolites on a high-resolution platform enables an automated selection of candidate molecules. The procedure consists of three major steps: (*i*) mass-based preselection, (*ii*) determination of pairwise similarity, and (*iii*) filtering according to empirically determined cutoffs (Figure 4.6).

Complete databases, such as Recon or HMDB were searched for molecules that correspond to the monoisotopic neutral mass of unknown metabolites, with Δmass ≤ 0.05. Molecules meeting the mass criterion were preselected and forwarded to the structural similarity check.

As second step, the structural similarity was determined for pairs of preselected candidate molecules and structurally known neighbors of the respective unknown metabolite. For each candidate, its maximum similarity values were maintained. Missing SDFs were retrieved just-in-time from PubChem by the function '`getIds`' of the R-package '*ChemmineR*' (version 2.30.0) [179].

To obtain final candidates, the set of preselected molecules was filtered according to cutoffs that were empirically determined during a 10-fold cross-validation. Candidates were selected, if at least one of their coefficients (fingerprint Tanimoto, MCS Tanimoto or MCS overlap) passes its cutoff criterion.

Cross-validation and optimization of parameters

As proof of principle and for parameter optimization a 10-fold cross-validation was performed across known molecules. Initial cutoffs WM were determined for fingerprint Tanimoto, MCS Tanimoto, and MCS overlap coefficients by the mean of each coefficient-distribution of neighboring versus random pairs of known metabolites weighted by their 25 or 75% quantiles (Equation 2.6, Figure 4.7).

$$WM = m_a + |m_a - q_a| \cdot \frac{m_b - m_a}{|m_b - q_b| + |m_a - q_a|}$$

$$\text{with} \quad q_x = \begin{cases} q_{[x,25\%]} & , m_x > min\,(m_a, m_b) \\ q_{[x,75\%]} & , else \end{cases} \quad , \text{ and} \quad x \in \{a, b\}$$

Equation 2.6: Weighted mean of distributions of coefficients
The mean of two coefficient distributions a and b is weighed by the 25 and 75% quantile. m_x and q_x refer to the mean and respective quantile of distribution a or b.

Function ‘`createFolds`’ of the R-package ‘*caret*’ version 6.0-78 [180] divided metabolites with known chemical structures into one training set and one test set, containing 90% and 10% of entries respectively. During each run of the cross-validation initial cutoffs were individually shifted by -0.5 to -5 in steps of 0.5. The final cutoff configuration was selected by the mean cutoffs that maximize the performance score, containing weighted sensitivity and specificity (Equation 2.7).

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP} \tag{a}$$

$$Performance = \frac{w \cdot Sensitivity + Specificity}{1 + w} \tag{b}$$

Equation 2.7: Statistical measure of performance
For measuring performance, weight w was set to 2 or 3 to prioritize sensitivity towards specificity. TP: true positive, TN: true negative, FP: false positive, FN: false negative.

Prioritizing sensitivity before specificity leads to a larger number of selected candidates (positives) and ensures that fewer correct candidates are missed. Indeed, a larger number of positives induces an increased rate of both, true positives (TPs) and true negatives (TNs).

2.2.10 Experimental verification of selected candidate molecules

To confirm automatically or manually selected candidate molecules, substances need to pass an experimental measurement (MS, LC-MS, or GC-MS) in a wet-laboratory. Candidate molecules can be substantiated, if candidates and respective unknown metabolites are not distinguishable by the suitable MS technique. If measurements of both aliquots show differences, candidates may be excluded. We sought experimental conformation of a selected set of unknown metabolites for which the most frequently observed reaction type, dehydrogenation reactions, were predicted. To verify or falsify these candidates of unknown metabolites we purchased all (6) corresponding molecules that were commercially available as pure substances (Table 2.1).

Candidate substances were dissolved in water at a concentration of 1 mg/mL by ultrasonification and, where appropriate, by addition of several droplets of methanol and diluted with the LC running solvent A to a concentration of 100 ng/mL. Analyses of candidate solutions were performed with LC-MS in negative ionization mode on a LTQ XL mass spectrometer (Thermo Fisher Scientific GmbH, Dreieich, Germany)

Candidate molecule	Supplier	Product number	Molar mass
9-octadecenedioic acid	Anward, Kowloon, Hong Kong	ANW-62167	312.23 g/mol
trans-2-nonenoic acid	Sigma-Aldrich Chemie GmbH, Steinheim, Germany	S354015	156.12 g/mol
3-nonenoic acid	Sigma-Aldrich Chemie GmbH	CDS000243	156.12 g/mol
8-nonenoic acid	Sigma-Aldrich Chemie GmbH	715433	156.12 g/mol
cis-9-tetradecenoic acid	Sigma-Aldrich Chemie GmbH	M3525	226.19 g/mol
trans-2-dodecenedioic acid	VWR International GmbH, Darmstadt, Germany	CAYM88820	228.14 g/mol

Table 2.1: Candidate molecules for experimental verification
Six commercially available candidate molecules for four unknown metabolites were forwarded to the experimental verification. Candidate molecules were selected according to their Δmass and prediction of dehydrogenation reaction (Δm and ΔRI).

coupled to a Waters Acquity UPLC system (Waters GmbH, Eschborn, Germany) at the Helmholtz Zentrum München. After sample injection the column (2.1 mm × 100 mm Waters BEH C18, 1.7 µm particle-size) was developed with a gradient of 99.5% solvent A (6.5 mM ammonium bicarbonate [pH 8.0]) to 98% solvent B (6.5 mM ammonium bicarbonate in 95% methanol). The flow rate was set to 350 µL/min for a run time of 11 minutes. The eluent was directly connected to the electrospray ionization source of the mass spectrometer.

For each candidate the pure substance and a spiked mixture with an extracted reference plasma sample was analyzed. For comparison the reference plasma containing the unknown compounds at natural abundance was measured as well. MS scans were recorded from 80 to 1000 m/z as well as data dependent MS/MS scans of the candidate masses.

We compared the RT of peaks in the EIC for the three measurements per metabolite. Especially the mixture of the pure substance and reference plasma should show just one peak, because otherwise both compounds could be separated during the chromatography and consequently are not identical. Finally, we checked if the MS^2 fragment spectra of the pure candidates and of the respective unknown metabolite in the matrix sample consisted of the same fragments with equal relative intensities. Candidates that could not be separated from the respective unknown metabolite by LC-MS analyses were labeled as verified.

2.2.11 Storage of predicted or verified metabolite annotations

Our automated workflow for annotation of unknown metabolites works best with data sets containing equally sized subsets of measured known and unknown metabolites. This is especially the case, if established non-targeted MS platforms (e.g. Metabolon$^{\mathrm{TM}}$) are used for quantification. To enable portability of annotations of unknown metabolites that have potentially been measured in several cohorts (and can be easily mapped), these predicted or verified information must be stored in a solid data structure. We designed a conceptual mySQL database consisting of two levels: (*i*) 'MeasuredMetabolite' and related tables contain information concerning measuring parameters (e.g. RI, mass-to-charge ratio, fragments, tissue, study, or platform), associations (e.g. GGM, mGWAS, or MWAS), annotations (e.g. predicted pathways, reactions, chemical formulas, or candidates), or identification (e.g. chemical structure and identification level) of (unknown) measured molecules; if (formerly unknown) measured metabolites are annotated or identified, they can be assigned to compounds of public database (*ii*) of table 'MetaboliteDetail' including various connected information (e.g. several database IDs, IUPAC name, synonyms, chemical formula and structure, monoisotopic and average mass, Simplified Molecular Input Line Entry System (SMILES), International Chemical Identifier (InChI), reactions, functional relations, and ontology information) (Figure 4.13). The suggested database helps to aggregate several annotations of overlapping sets of unknown metabolites that have been collected independently, or to extract existing annotations for a given set of unknown metabolites occurring in results of novel experiments.

CHAPTER 3

Results and discussion I: Characterization of PC compositions measured as lipid sums

Labels of some metabolites in targeted as well as in non-targeted mass spectrometry (MS) indicate ambiguous molecular structures. If the applied analytical method cannot separate compounds that share equal chemical or physical properties, metabolite sums instead of single compounds are measured.

With the contents of this chapter (except the imputation of metabolite concentrations) including respective materials and methods I prepared a manuscript that I will submit for publication concurrently; Quell et al. [149].

3.1 Background

Phosphatidylcholines (PCs) are among the major constituents of cell membranes [129, 130], important compounds of lipoproteins [131], and crucial in the energy metabolism. PCs consist of a polar head group with phosphocholine and a nonpolar part with two fatty acid chains (with different lengths and desaturation) connected to the *sn*-1 or *sn*-2 positions of the glycerol backbone via an ester (acyl) or ether (alkyl) bond [132–134]. Besides the abundance of total PC, slight changes in the chemical structure of PCs imply large functional differences. As an example, stable omega-6 to omega-3 free fatty acid ratios of 4 to 1 are associated with a 70% lower total mortality of patients suffering from cardiovascular diseases [135]. When fatty acids are bond to PCs, (besides their chain lengths and desaturation,) the fluidity, for instance, of PC 16:1/16:1 (fatty acyl/alkyl position level; cf. Figure 3.1) is higher than the fluidity of PC 16:0/16:1 or PC 16:0/16:0 [136]. Even for PCs with a fixed desaturation, e.g. PC 18:0/18:1, the fluidity is maximized when double bonds are in the delta-9 position [137]. Therefore, meaningful biochemical interpretations need to be as close as possible to the actual chemical structure of lipids.

Today, the availability of modern high-throughput lipidomics platforms allows quantification of numerous lipids in thousands of blood samples of epidemiological cohorts. In particular, the MS-based targeted Absolute*IDQ*$^{\mathrm{TM}}$ kit has produced large data sets of metabolites, such as amino acids, acylcarnitines, PCs, lysoPCs, and SMs [26, 27, 141, 143, 144, 181–184]. Absolute*IDQ*$^{\mathrm{TM}}$ measures PCs during multiple reaction monitoring (MRM) scans by selecting the total m/z (Q1) and the head group (Q3) (example in Table 3.1). As a consequence, the total length of both fatty acid chains and number of double bonds is determined, but their division and configuration at the *sn*-1 and *sn*-2 positions remain unspecified and metabolites are annotated on the lipid species level only (Figure 3.1). Furthermore, PC and isobaric PC-O, in which one fatty acid chain is longer by one and attached by an alkyl bond, are not analytically separated (e.g. PC 33:1 and PC O-34:1 in Table 3.1). In various large-scale mGWAS and MWAS, multiple of these PC measures have been reported to associate with common genetic variants and various diseases such as Alzheimer's disease [142, 143], Type 2 Diabetes [141], coronary artery disease [144], or gastric cancer [145].

lipid class level mass		**PC (745)**		
lipid species level		**PC 33:1**		**PC O-34:1**
fatty acid (acyl/alkyl) level		**PC 15:0_18:1**	**PC 17:0_16:1**	**PC O-16:0_18:1**
Absolute*IDQ*$^{\mathrm{TM}}$	Q1 (total m/z)*	746.6	746.6	746.6
	Q3 (head group)	184	184	184
Lipidyzer$^{\mathrm{TM}}$	Q1 (total m/z)*	804.6	804.6	804.6
	Q3 (1 fatty acid chain)	241.2	269.2	—

Table 3.1: Multiple reaction monitoring leading to different precision levels
As an example, differently set MRMs are demonstrated on selected isobaric compounds of PC (745). While Absolute*IDQ*$^{\mathrm{TM}}$ measures the total m/z (Q1) and the head group (Q3) during MRM scans and cannot separate selected isobaric compounds (lipid species level), Lipidyzer$^{\mathrm{TM}}$ targets one of both fatty acid chains (Q3) beside the total m/z (fatty acid level). Thus, both chain lengths including their desaturation are determined by Lipdyzer$^{\mathrm{TM}}$. To isolate PC aa from the isobaric PC ae, Lipidyzer$^{\mathrm{TM}}$ preferably selects the expected mass of the odd-chain fatty acid (if applicable; but does not measure PC aes). All masses are given in Dalton. *) Absolute*IDQ*$^{\mathrm{TM}}$ measures PCs in positive mode ($+[\mathrm{H}]^+ \widehat{=} 1$ Da); Lipidyzer$^{\mathrm{TM}}$ measures PCs in negative mode ($+[\mathrm{M}+\mathrm{OAc}]^- \widehat{=} 49$ Da).

The gap between the lipid species level and the concrete chemical structure needs

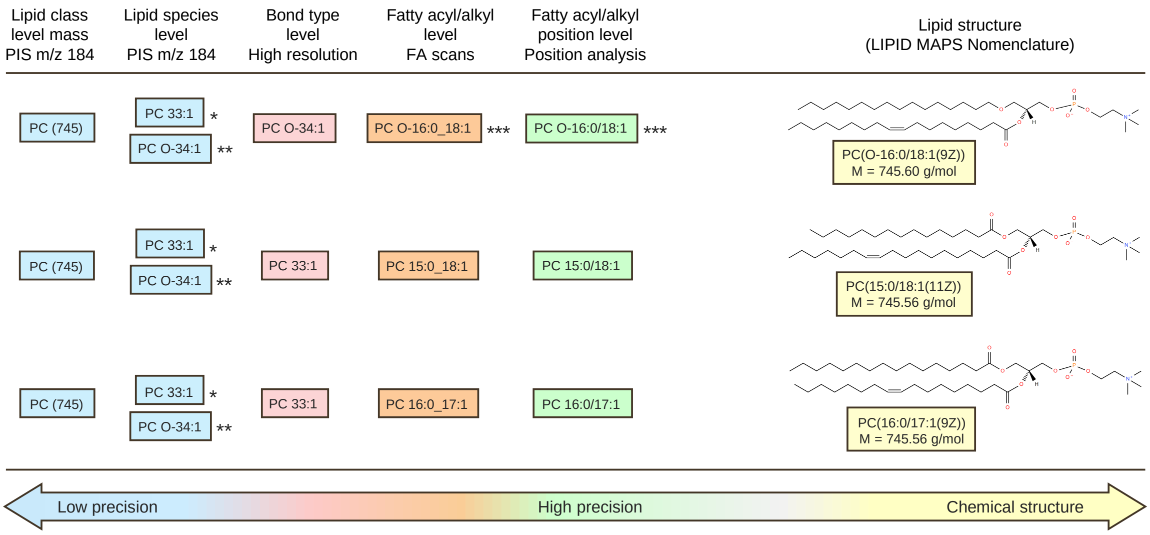


Figure 3.1: Platform precision levels for notation and measurement
Fatty acids can be specified on different precision levels ranging from the lipid class level (low precision) to the lipid structure (maximum precision). Current mass spectrometry approaches separate metabolites on the lipid species, or fatty acyl/ alkyl position level. Adapted figure from Liebisch et al. by permission from the Journal of Lipid Research [162], copyright 2013. Chemical structures were taken from Lipid Maps: LMGP01020003, LMGP01010541, LMGP01010571 [175].

to be reduced to use results based on these measured metabolites reasonably, which is currently only possible under the following assumptions: even chain lengths are more abundant than odd chain lengths [138] (e.g. measurement of PC (745) is annotated as PC O-34:1, in which one of the fatty acids is attached by an alkyl bound, instead of the isobaric PC 33:1); especially C16:0, C18:0, C18:1, C18:2, and C20:4 fatty acids are very common in human plasma [139] (e.g. leading to label PC O-16:0_18:1); and further more specific expectations regarding their concrete configuration [140, 141] (e.g. ordering of fatty acids to *sn*-1 and *sn*-2 positions, PC O-16:0/18:1) (Figure 3.1).

Recently published experimental non-targeted one or two-dimensional LC-MS technologies study retention behavior and fragment spectra of complex lipids to discriminate complex lipids on the more precise fatty acyl/alkyl level, in which the length and desaturation of both fatty acid chains is determined, but their ordering (*sn*-1 or *sn*-2) and position of double bonds remains unspecified (Figure 3.1). Since current established lipidomics platforms (such as AbsoluteIDQ^{TM}) generate measurements on the lipid species level, these studies provide annotations on both precision levels. However, they cannot specify quantitative compositions or main constituents of metabolites measured on the coarser lipid species level in actual biological samples. [185–187]

Developed complex lipidomics FIA platforms, such as Lipidyzer$^{\text{TM}}$, perform standardized targeted high-throughput measurements by an ion trap that targets the total m/z (Q1) and one of the two fatty acid chains (Q3) during MRM scans (example in Table 3.1). Thus, measurements meet the fatty acyl/alkyl level. To improve the readability, I use the term fatty acid level instead of fatty acyl/alkyl level in this work. Although annotations on the fatty acid level are far from the specific molecular structure, they already allow more precise interpretations by a significantly reduced number of possible constituents.

To prevent remeasurements of samples that have already been quantified on the lipid species level (that would be unnecessarily expensive, or impossible if samples are already used up or reserved for further analyses), their composition should be determined on the more accurate fatty acid level. Especially for metabolites measured non-targeted LC-MS devices, there are decision rule based (LipidMatch [188] and Lipid Data Analyzer (LDA) [189]) or comparative (Greazy [190] and MS-DIAL [191]) approaches that use retention times and fragment spectra to predict complex lipids. However, these methods cannot be applied to metabolites of the MRM based FIA platform Absolute*IDQ*$^{\text{TM}}$. Therefore, I compare metabolite levels measured by Absolute*IDQ*$^{\text{TM}}$ and Lipidyzer$^{\text{TM}}$ in the same samples of a cohort. In former metabolomics studies data of several complementary platforms has been combined to increase the number of metabolites or incorporate metabolites covering multiple pathways and molecular classes [181, 184]. Platform comparisons have successfully been applied to find the most suitable analytical technique for exemplary use cases [192], and to compare metabolite levels (e.g. for replication of scientific results) directly or especially in case of non-targeted measurements, based on genetic or inter-metabolome associations [181, 193]. In this work, I perform a comparison of metabolite levels that were measured by two platforms covering these precision levels to estimate quantitative compositions shifting metabolites from the lipid species to the fatty acid level.

Overview of the PC characterization approach introduced

In this study, I introduced an automatic pipeline to characterize ambiguously labeled PC sums that were measured on the lipid species level by identifying their main constituents of the fatty acid level and specifying their quantitative contributions. To this end, I used 76 and 109 PCs, respectively, that were quantified on the targeted lipidomics platforms Absolute*IDQ*$^{\text{TM}}$ and Lipidyzer$^{\text{TM}}$ in 223 human plasma samples of healthy young men. In a first step, I assigned all theoretically possible PCs (fatty acid level) to the 76 PC sums by systematically distributing the total length of their fatty acid chains and desaturation to the *sn*-1 and *sn*-2 positions. The resulting quantitative descriptions of PC compositions were filtered by measured molecules with at least 25% coverage to prepare concentration-based quantitative estimations. Quantitative PC compositions and imputation of PC concentrations (fatty acid level) were replicated in an independent cohort to demonstrate their stability. In a last step, I estimated the fraction of PC sums that could not be mapped to any measured PC of the fatty acid level.

3.2 Qualitative composition of PC sums

To map PCs measured on the fatty acid level to PCs measured on the lipid species level, we systematically distributed the total numbers of carbon atoms and double bonds to two fatty acid chains and collected all possible isobars (Table 3.2). We gathered 150 to 456 theoretically possible isobaric PCs of the fatty acid level for 76 PC

PC species	Isobaric PC species	Example	Change in sum formula	Change in mass
PC $x : y$		PC 32:0		
	PC $x + 1 : y + 7$	PC 33:7	$+CH_2$ −14 H	− 0.093 900 Da
	PC O-$x + 1 : y$	PC O-33:0	$+CH_2$ +2H −O	+ 0.036 385 Da
	PC O-$x + 2 : y + 7$	PC O-34:7	+2(CH_2) −12H −O	− 0.057 515 Da
	[$^{13}C_1$]SM $x + 4 : y$	[$^{13}C_1$]SM 36:0	+^{13}C +5H +N −2O	+ 0.055 724 Da
PC O-$x : y$		PC O-32:0		
	PC O-$x + 1 : y + 7$	PC O-33:7	$+CH_2$ −14H	− 0.093 900 Da
	PC $x - 1 : y$	PC 31:0	+O $-CH_2$ −2H	− 0.036 385 Da
	PC $x : y + 7$	PC 32:7	+O −16H	− 0.130 285 Da
	[$^{13}C_1$]SM $x + 3 : y$	[$^{13}C_1$]SM 35:0	+N +H +^{13}C −^{12}C −O	+ 0.019 339 Da

Table 3.2: Isobaric phosphatidylcholine sums within mass resolution of the MS device

MS platforms that quantify metabolites on the lipid species level cannot separate isobaric PC species. Furthermore, fatty acid chain lengths x and double bounds y can be systematically distributed to the *sn*-1 and *sn*-2 position to receive theoretically possible isobaric PCs on the fatty acid level. Changes in mass have been calculated according to the changes in sum formulas and monoisotopic atom weights w: w(H) = 1.007 825 Da, w(C) = 12.000 000 Da, w(^{13}C) = 13.003 355 Da, w(O) = 15.994 915 Da, w(N) = 14.003 074 Da. This table is part of my prepared manuscript as well; Quell et al. [149].

sums measured on the lipid species level (Supplemental Table A.1). Biocrates$^{\text{TM}}$ provides a list with the most probable 2 to 17 PCs (fatty acid level) per PC sum that are tagged in the supplemental table [194]. In addition, we highlighted 93 of 109 compounds in this table that have been quantified on the Lipidyzer$^{\text{TM}}$ platform in our data set. The isobaric [^{13}C$_1$]SM $x + 4 : y$ was added as a possible compound, if it has not been quantified with the kit and, thus, has not been subtracted from the PC concentration in the isotope correction step of the Absolute*IDQ*$^{\text{TM}}$ processing. This was the case for 33 PC aa C$x : y$ and 38 PC ae C$x + 1 : y$ species.

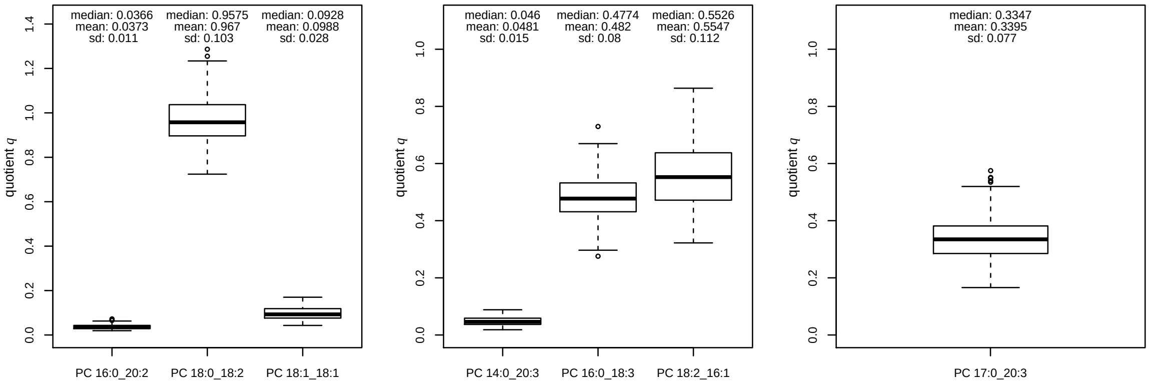
3.3 Quantitative composition of PC sums

For 38 (25 PC aa and 13 PC ae) out of the 68 PC sums (that were quantified with the Absolute*IDQ*$^{\text{TM}}$ and passed the QC), in total, 69 (of 93 mapped) respective PCs of the fatty acid level (Lipidyzer$^{\text{TM}}$) were available at sufficient coverage (25%) in the HuMet Study [25] to estimate quantitative compositions. Each PC sum (lipid species level) consists of one to four measured PCs of the fatty acid level, e.g.: PC aa C36:2 = PC 16:0_20:2 + PC 18:0_18:2 + PC 18:1_18:1 + R, PC aa C34:3 = PC 14:0_20:3 + PC 16:0_18:3 + PC 18:2_16:1 + R, or PC aa C32:2 = PC 14:0_18:2 + R (Supplemental Table A.2). Here, R corresponds to the sum of respective non-measured PCs of the fatty acid level that are specified in Supplemental Table A.1. Consistent with the resolution of the MS techniques, we use the fatty acid level of the short hand notation of lipids for PCs, e.g. PC $x_1 : y_1\_x_2 : y_2$ (Figure 3.1) [162]. PCs measured on the lipid species level are labeled according to the Biocrates$^{\text{TM}}$-specific notation, e.g. PC aa C$x : y$ or PC ae C$x : y$, corresponding to PC $x : y$ or PC O-$x : y$. Notably, PC aa or PC ae labels were chosen under the assumption of even chain lengths of fatty acids.

3.3.1 Estimation of factors f

In a first step, we calculated factors f (mean of divided concentrations $\frac{PC_{\text{fatty acid level}}}{PC_{\text{lipid species level}}}$) estimating the proportion of each measured PC of the fatty acid level based on the concentration of the respective PC sum (lipid species level) in the 223 HuMet samples (Table 3.3). As an example, three measured PCs of the fatty acid level, namely PC 16:0_20:2, PC 18:0_18:2, PC 18:1_18:1, are isobaric constituents of the PC aa C36:2 measured at the lipid species level. The factors f [5% and 95% confidence intervals] for these three species are 0.037 [0.023, 0.058], 0.967 [0.817, 1.14] and 0.099

[0.060, 0.147], respectively (Figure 3.2(a)). For 27 out of the 38 investigated PC sums, we identified at least one PC measured on the fatty acid level with a factor greater than 20%. Three PC sums were characterized by at least two PCs measured on the fatty acid level with $f > 0.2$. As an example, PC aa C34:3 consists of PC 14:0_20:3 (0.048 [0.025, 0.073]), PC 16:0_18:3 (0.482 [0.354, 0.617]), and PC 18:2_16:1 (0.555 [0.381, 0.741]) (Figure 3.2(b)). In ten cases, we found one major component with a proportion greater than 80%. Interestingly, one PC sum contains a PC of the fatty acid level with an odd numbered fatty acid chain: PC 17:0_20:3 is a major component of PC ae C38:3 ($f = 0.34$) (Figure 3.2(c)). For 11 PC sums, factors f for all measured constituents sum up to 80 − 120%. In four cases the sum of factors f is larger than 1.2.



(a) PC aa C36:2 (b) PC aa C34:3 (c) PC aa C38:3

Figure 3.2: Quantitative compositions of PCs measured on the lipid species level and respective PCs measured on the the fatty acid level

The mean values of quotients of PCs measured on the fatty acid level and the respective PC sums (lipid species level) estimate factors f (factor f is the mean of quotients q over all samples). Concentrations of PCs measured on the lipid species level (i) can be multiplied with f_{ij} to impute the concentrations of PCs on the fatty acid level (j). This figure is part of my prepared manuscript as well; Quell et al. [149].

3.3.2 Variation of factors f

Since factors f might vary, we were interested in differences of their distributions within our data set. To this end, we examined changes in factors f across subjects and during lipid-related metabolic challenges such as 36h fasting (FASTING), physical activity test (PAT), and after ingestion of a lipid-enriched meal (OLTT). For testing differences between subjects, we compared the distributions of quotients q of 56 time points of each of our four subjects (factors f correspond to the mean

Lipid species	Absolute*IDQ*™	Lipidyzer™	R² of	Fraction	Factor	Conf. interval		Cat.	Stab.	Stab.
(neutral mass)	Metabolite	Metabolite	LM [%]	[%]	*f*	5%	95%	*f*	*f* | ic	Subj.
PC 30:0 (705)	PC aa C30:0	PC 16:0_14:0	83.3	81.7	0.3584	0.2248	0.5098	I		Y
		PC 18:0_12:0		18.3	0.0820	0.0462	0.1495		y | y	
PC 32:0 (733)	PC aa C32:0	PC 16:0_16:0	35.3	97.7	1.0218	0.7561	1.3897	II°		
		PC 18:0_14:0		2.3	0.0231	0.0138	0.0347			y
PC 32:1 (731)	PC aa C32:1	PC 14:0_18:1	95.1	27.8	0.2981	0.1773	0.5843	I°	| y	
		PC 16:0_16:1		72.2	0.7669	0.5433	0.9891	I	| Y	
PC 32:2 (729)	PC aa C32:2	PC 14:0_18:2	78.9	100	1.2713	0.8229	1.7371	II$^+$		
PC 34:1 (759)	PC aa C34:1	PC 16:0_18:1	86.6	99.5	1.4159	1.1793	1.6821	II$^+$	| Y	Y
		PC 18:0_16:1		0.4	0.0057	0.0028	0.0120		| y	
		PC 20:0_14:1		0.1	0.0014	0.0009	0.0022			
PC 34:2 (757)	PC aa C34:2	PC 14:0_20:2	49.4	< 0.05	0.0006	0.0004	0.0010	$^+$	| y	y
		PC 16:0_18:2		98.1	1.5482	1.2968	1.8214	II	| Y	Y
		PC 18:1_16:1		1.9	0.0247	0.0155	0.0371		y | y	
PC 34:3 (755)	PC aa C34:3	PC 14:0_20:3	85.4	4.5	0.0481	0.0252	0.0729	°		
		PC 16:0_18:3		44.7	0.4820	0.3539	0.6165	I	Y | Y	
		PC 18:2_16:1		50.9	0.5547	0.3811	0.7406	I		
PC 34:4 (753)	PC aa C34:4	PC 14:0_20:4	53.6	100	0.3681	0.2438	0.5650	I		
PC 36:0 (789)	PC aa C36:0	PC 18:0_18:0	14.3	100	0.4844	0.3066	0.7470	I		Y
PC 36:1 (787)	PC aa C36:1	PC 16:0_20:1	79.0	3.5	0.0319	0.0231	0.0447	°	| y	
		PC 18:0_18:1		96.5	0.8819	0.7171	1.0794	II	Y | Y	Y
PC 36:2 (785)	PC aa C36:2	PC 16:0_20:2	57.9	3.4	0.0373	0.0226	0.0581	°	y | y	
		PC 18:0_18:2		87.7	0.9670	0.8172	1.1362	II	Y | Y	Y
		PC 18:1_18:1		8.9	0.0988	0.0601	0.1467			
PC 36:3 (783)	PC aa C36:3	PC 16:0_20:3	80.4	53.6	0.6060	0.3434	0.8409	I°	Y | Y	
		PC 18:0_18:3		1.5	0.0175	0.0100	0.0279		| y	
		PC 18:1_18:2		44.8	0.5078	0.3177	0.7643	I	Y |	
PC 36:4 (781)	PC aa C36:4	PC 14:0_22:4	65.2	0.4	0.0016	0.0011	0.0025			
		PC 16:0_20:4		78.2	0.2943	0.2393	0.3555	I		
		PC 18:1_18:3		2.5	0.0109	0.0033	0.0220			
		PC 18:2_18:2		18.8	0.0774	0.0354	0.1558		| y	
PC 36:5 (779)	PC aa C36:5	PC 14:0_22:5	17.2	39.0	0.0172	0.0101	0.0262			
		PC 18:2_18:3		61.0	0.0310	0.0113	0.0615			y
PC 36:6 (777)	PC aa C36:6	PC 14:0_22:6	32.3	100	0.8810	0.4947	1.2925	II°		Y
PC 38:0 (817)	PC aa C38:0	PC 18:0_20:0	34.5	100	0.1629	0.1193	0.2153			y
PC 38:3 (811)	PC aa C38:3	PC 18:0_20:3	80.7	94.3	0.6485	0.4476	0.8496	I	Y | Y	
		PC 18:1_20:2		2.9	0.0197	0.0126	0.0299			
		PC 18:2_20:1		2.7	0.0172	0.0070	0.0436			
PC 38:4 (809)	PC aa C38:4	PC 16:0_22:4	58.8	18.9	0.0758	0.0504	0.1059		y | y	
		PC 18:0_20:4		59.5	0.2343	0.1971	0.2867	I		Y
		PC 18:1_20:3		19.9	0.0808	0.0490	0.1500			
		PC 18:2_20:2		1.6	0.0065	0.0028	0.0126		y |	
PC 38:5 (807)	PC aa C38:5	PC 16:0_22:5	67.7	80.5	0.5515	0.4380	0.6725	I	Y | Y	
		PC 18:1_20:4		14.4	0.0983	0.0687	0.1286			y
		PC 18:2_20:3		5.1	0.0349	0.0201	0.0647			y

Table continued (p. 2/2)

Lipid species	**Absolute*IDQ*™**	**Lipidyzer™**	**R^2 of**	**Fraction**	**Factor**	**Conf. interval**		**Cat.**	**Repl.**	**Stab.**
(neutral mass)	**Metabolite**	**Metabolite**	**LM [%]**	**[%]**	f	**5%**	**95%**	f	f **| ic**	**Subj.**
PC 38:6 (805)	PC aa C38:6	PC 16:0_22:6	77.1	98.5	1.2431	1.0168	1.5009	II$^+$	Y | Y	Y
		PC 18:2_20:4		1.5	0.0190	0.0118	0.0279			
PC 40:3 (839)	PC aa C40:3	PC 20:0_20:3	0.2	100	0.4892	0.1980	0.8453	I		
PC 40:4 (837)	PC aa C40:4	PC 18:0_22:4	67.2	88.0	0.7065	0.4936	0.9503	I$^\circ$	| Y	
		PC 20:0_20:4		12.0	0.0952	0.0644	0.1461			y
PC 40:5 (835)	PC aa C40:5	PC 18:0_22:5	76.5	93.0	0.7863	0.6343	0.9819	I$^\circ$		
		PC 18:1_22:4		7.0	0.0588	0.0373	0.0840		y | y	
PC 40:6 (833)	PC aa C40:6	PC 18:0_22:6	82.7	91.8	1.0540	0.8588	1.2734	II$^\circ$	| Y	Y
		PC 18:1_22:5		7.0	0.0718	0.0429	0.1039			
		PC 18:2_22:4		1.1	0.0126	0.0076	0.0186			y
PC 42:6 (861)	PC aa C42:6	PC 20:0_22:6	5.9	100	0.8482	0.5587	1.2468	II$^\circ$		
PC 33:1 (745)	PC ae C34:1	PC 15:0_18:1	57.8	69.7	0.0988	0.0705	0.1373			y
		PC 17:0_16:1		30.3	0.0438	0.0307	0.0601			y
PC 33:2 (743)	PC ae C34:2	PC 15:0_18:2	7.1	100	0.1059	0.0698	0.1541			
PC 35:1 (773)	PC ae C36:1	PC 17:0_18:1	64.4	100	0.2147	0.1601	0.2878	I	| Y	Y
PC 35:2 (771)	PC ae C36:2	PC 17:0_18:2	52.8	100	0.2275	0.1837	0.2799	I		Y
PC 35:3 (769)	PC ae C36:3	PC 15:0_20:3	49.9	100	0.0700	0.0458	0.0950			
PC 35:4 (767)	PC ae C36:4	PC 15:0_20:4	26.0	100	0.0526	0.0339	0.0793		y | y	
PC 37:3 (797)	PC ae C38:3	PC 17:0_20:3	64.5	100	0.3395	0.2248	0.4767	I	Y | Y	
PC 37:4 (795)	PC ae C38:4	PC 17:0_20:4	64.1	100	0.1978	0.1495	0.2547		y | y	y
PC 37:5 (793)	PC ae C38:5	PC 17:0_20:5	2.2	100	0.0263	0.0145	0.0408			
PC 37:6 (791)	PC ae C38:6	PC 15:0_22:6	12.6	100	0.0677	0.0433	0.1084		y | y	
PC 39:1 (829)	PC ae C40:1	PC 18:2_22:6	20.3	100	0.5664	0.3782	0.8648	I		
PC 39:5 (821)	PC ae C40:5	PC 17:0_22:5	19.7	100	0.1234	0.0708	0.1735		y | y	y
PC 39:6 (819)	PC ae C40:6	PC 17:0_22:6	56.3	100	0.1720	0.1216	0.2262		y | y	y

Table 3.3: Quantitative composition of phosphatidylcholines
Each PC sum measured on the lipid species level (Absolute*IDQ*$^{\text{TM}}$) consists of a composition of PCs measured on the fatty acid level (Lipidyzer$^{\text{TM}}$). As Absolute*IDQ*$^{\text{TM}}$ cannot distinguish between acyl and alkyl bound types, the isobaric PC (lipid species) followed by the lipid class level mass is used. A linear model estimated R^2 as the percentage of variance of the PC sum that can be explained by the variances of the PCs of the fatty acid level. In contrast to factor f, fraction is the quantitative ratio of PCs of the fatty acid level within the measured compounds and sums up to 100% (per PC composition). The measured concentrations of PC sums can be multiplied by the factor f to impute the concentrations of respective PCs of the fatty acid level. The percentiles represent (5% and 95%) confidence intervals of this factor. Categories 'I' and 'II' correspond to factors f larger than 0.2 or larger than 0.8 respectively. The symbol '$^\circ$' was added, if aggregated factors f per PC sum are between 0.9 and 1.1, or '$^+$' if factors f are larger than 1.1. Factors f (or imputed concentrations (ic)) have been replicated, if distributions of quotients q in HuMet and in a subset of QMDiab, containing healthy men aged $\leq$40y, (or distributions of observed and imputed concentrations of PCs of the fatty acid level in the subset of QMDiab) were not significantly different ($\alpha = 0.01$). The stability of the factor f within subjects of HuMet was calculated by an ANOVA (distributions of quotients q were not significantly different according to an α level of 0.01). Capital letters indicate that replication or stability has been shown for the main compound (constituent with largest f) of a PC composition. A similar version of this table is part of my prepared manuscript as well; Quell et al. [149].

value of quotients q across all subjects and time points, cf. Equation 2.1). In 18 of 38 PC sums, distributions of quotients of the constituent with the largest f are not significantly different (Supplemental Table A.3) between subjects. Variations of f within challenges appear to be smaller compared to variations across subjects (average: $SD_{\text{subjects}} = 0.100$, $SD_{\text{fasting}} = 0.067$, $SD_{\text{sport}} = 0.043$, $SD_{\text{OLTT}} = 0.058$, and $SD_{\text{all time points}} = 0.050$; Supplemental Figure A.1). Since only four data points are available per baseline and intervention time point in each challenge, the statistical power is too low to identify significant differences in f depending on the status of lipid metabolism (Supplemental Table A.3).

3.4 Replication of quantitative compositions in another cohort

To test the transferability of factors f calculated from the HuMet study focusing on healthy male subjects (aged $22-33$y) of European ethnicity to other cohorts, we analyzed data from the two metabolomics platforms for participants of QMDiab with 151 healthy controls and 154 patients with diabetes (aged $17-81$y).

3.4.1 Comparison of factors f between data sets

First, we sought replication of the main constituents (largest f) of the 38 PC sums in the 31 healthy male participants of QMDiab aged below 40y. For 37 PC sums, the same PCs measured on the fatty acid level were identified as the main constituents with the largest f. One PC sum was not reported in QMDiab. For 13 PC sums, main factors f showed no significant difference between pairwise distributions of q in the two cohorts (Table 3.3, Supplemental Figure A.2 and Supplemental Table A.4). In general, factors f tend to be higher in samples of QMDiab, compared to factors f in samples of HuMet (Figure 3.3). Largest differences between factors f in both cohorts were identified for PCs of the fatty acid level containing C20:4 except for two PCs with C20:4 and a second fatty acid chain with an odd number of C-atoms (PC 15:0_20:4, PC 17:0_20:4).

QMDiab also allows calculation of factors f based on heterogeneous subsets including samples of both genders and patients suffering from type two diabetes. We compared the distributions of quotients q of three groups containing male controls, female controls, and cases, to estimate their variation: in 33 of 37 PC sums, factors f of the main compounds are not significantly different for these subsets. In the remaining four cases, the main factor is below 0.001 or not available (due to $> 75\%$ missing concentrations).

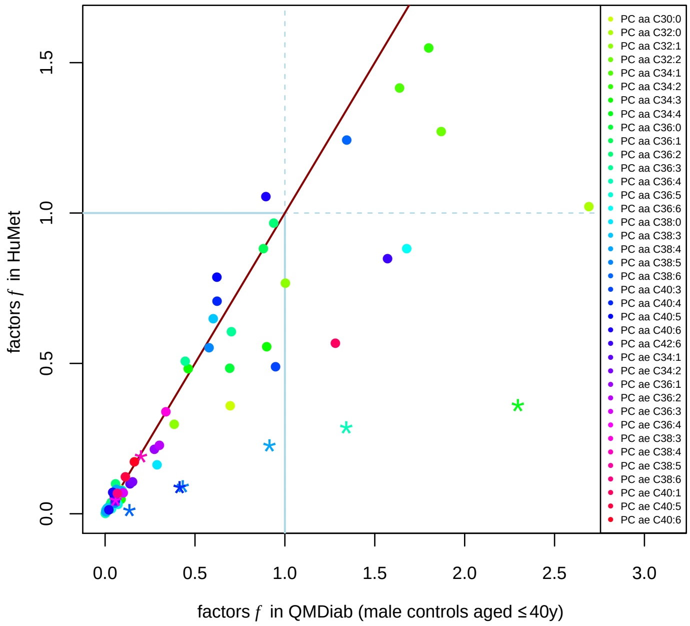


Figure 3.3: Factors f in HuMet and QMDiab
In Average, factors f tend to be slightly higher in a subset of QMDiab (male controls aged ≤ 40y) compared to factors f in HuMet. The red line indicates expected values: pairs of factors f that are equal in both cohorts. The blue lines show the theoretical maximum of factors f (otherwise the constituents have a higher concentration than the respective lipid species). PCs of the fatty acid level that contain C20:4 as one of their lipid side chains are marked with an asterisk. This figure is part of my prepared manuscript as well; Quell et al. [149].

3.4.2 Imputation of PC concentrations across data sets

Furthermore, we tested the transferability of factors f from HuMet to a subset of 31 healthy male participants aged ≤ 40y of QMDiab on concentration level (Figure 3.4). Therefore, we multiplied PC concentrations (lipid species level) of the subset of QMDiab by respective factors f that were estimated in HuMet, to impute PC concentrations (fatty acid level) and subsequently compared imputed to measured concentrations. 19 of 38 PC compositions showed no significant difference between distributions of measured and imputed concentrations of the main constituents (Table 3.3, Supplemental Table A.4, and Supplemental Figure A.3). Median concentrations of another four main constituents were within the $5-95\%$ confidence interval of factors f. In the remaining 15 PC compositions, imputed concentrations are smaller than observed (measured) concentrations.

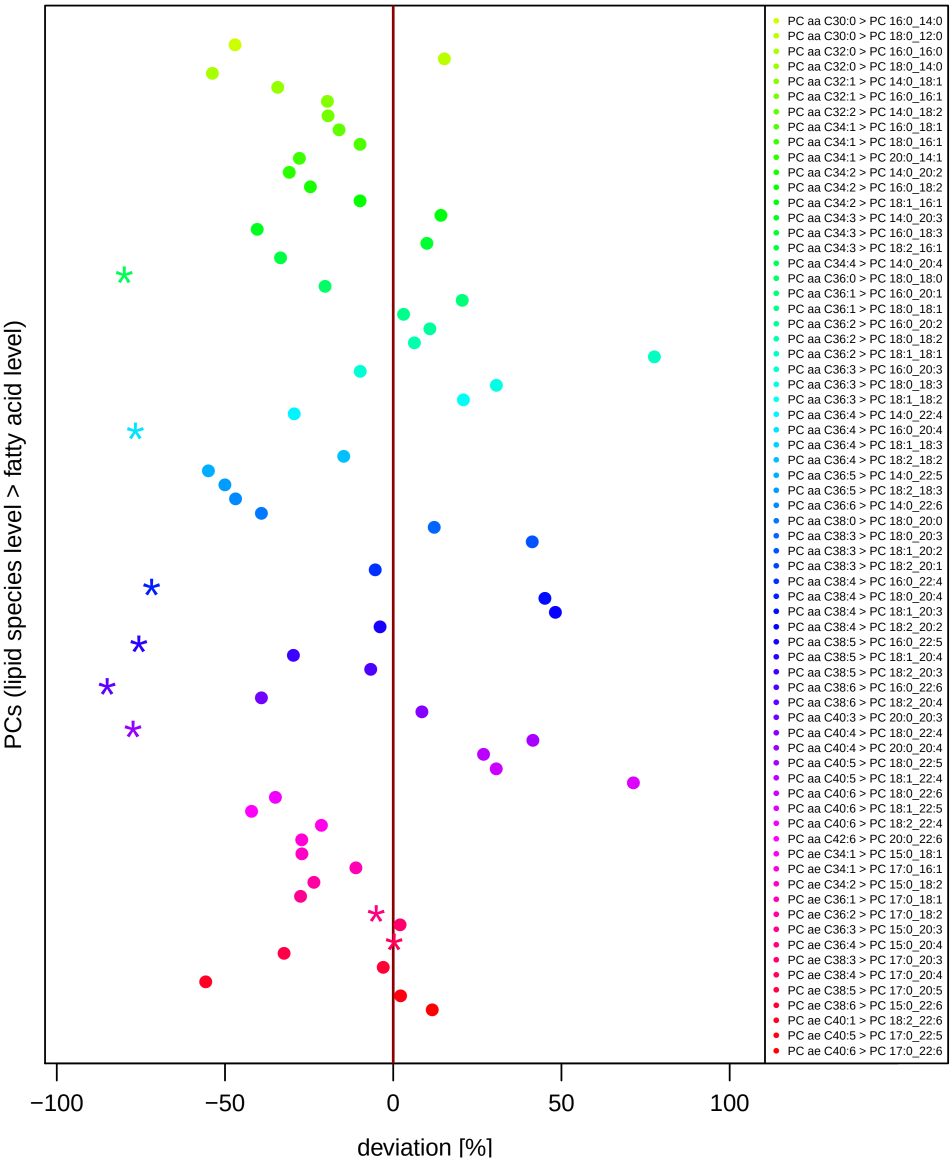
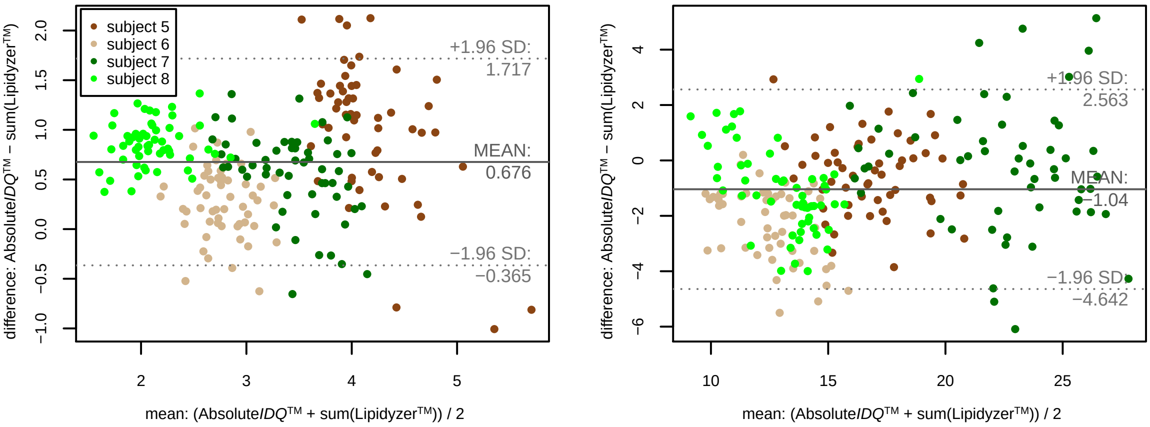


Figure 3.4: Relative deviation of measured vs. imputed median metabolite levels PC concentrations measured on the lipid species level in QMDiab were multiplied by factors f estimated in HuMet to impute levels on the fatty acid level. In compositions that are close to the red line, imputed values correspond to measured values. *) PCs cont. C20:4.

3.5 Contribution of non-measured constituents in PC sums

Out of 150 to 456 theoretically possible PCs of the fatty acid level per PC sum, the platform comparison only allowed mapping of one to four of these PCs to the respective sums to determine factors f (Supplemental Table A.1). To estimate the fraction of each PC sum that could be explained by measured PCs of the fatty acid level in our approach, we generated linear regression models (estimation of the explained variance; e.g. PC aa C36:2 $\sim$ $b\cdot$ (PC 16:0_20:2 + PC 18:0_18:2 + PC 18:1_18:1)) and Bland-Altman plots (comparison of absolute concentrations between both platforms; Figure 3.5) for the 38 PCs listed in Table 3.3.



(a) PC aa 40:4 = PC 18:0_22:4 + PC 20:0_20:4

(b) PC aa 34:3 = PC 14:0_20:3 + PC 16:0_18:3 + PC 18:2_16:1

Figure 3.5: Deviation and mean of PCs measured on different platforms
Shifts in difference (possibly dependent on the mean) of concentrations of PC sums (lipid species level) and respective constituents (fatty acid level) indicate whether both techniques measure the same compounds in a comparable accuracy (Bland-Altman plot). Colors refer to four subjects. In 19 of 25 cases the mean difference is within 2 SD of the concentration of the PC measured on the lipid species level. A similar version of this figure is part of my prepared manuscript as well; Quell et al. [149].

In median, 58% (mean: 52%, SD: 28%) of the variance of PC sums can be explained by the variances of measured PCs of the fatty acid level (Table 3.3 and Supplemental Table A.5). In consequence, 42% of the variance of PC sums is related to further constituents or other factors such as experimental variation.

We investigated systematic differences in metabolite concentrations of both measuring techniques under the assumption that all relevant constituents have been measured (Supplemental Table A.2). Since both platforms reported metabolite levels in the same quantitative units (µmol/L), we compared concentrations of PC sums (lipid species level) to concentrations of mapped PCs of the fatty acid level (e.g. *conc.*(PC aa C36:2) − $\Big($*conc.*(PC 16:0_20:2) + *conc.*(PC 18:0_18:2) + *conc.*(PC 18:1_18:1)$\Big)$) and the average of both concentrations (Bland-Altman plots). In 19 of 25 PC aa compositions, the mean of this difference is within 2 SD of the PC sum (in 12 of those compositions mean difference is within 1 SD). Small mean differences on the y-axis (in relation to absolute concentrations) and small variation (e.g. PC aa C40:4 = PC 18:0_22:4 + PC 20:0_20:4, absolute mean difference: 0.68; mean difference is within 1 SD of the PC sum) indicate concurrent measurements of both platforms (Figure 3.5(a)). In 10 cases (seven cases within 1 standard deviation (SD) and further two cases within 2 SD), in which mean differences are negative, concentrations of PCs measured on the fatty acid level are larger than those of PCs measured on the lipid species level (Figure 3.5(b)). All 13 compositions of PC ae sums show positive mean deviations larger than two. Mean concentrations on the x-axis of Bland-Altman plots enable identification of trends in differences of measuring methods that depend on absolute values (e.g. PC aa C34:4, PC aa C36:4, PC aa C36:5, PC aa C38:0).

3.6 Discussion

Absolute*IDQ*$^{\text{TM}}$ has emerged as an established measurement kit for quantification of lipids, such as phosphatidylcholines (PCs), in blood from thousands of participants in large epidemiological cohorts [26, 27, 141, 144, 145, 181, 183, 184, 195, 196] and in projects using various matrices of human, animal, or plant samples [197–201]. As the kit quantifies PCs on the level of lipid species, each PC sum measure can, in theory, consist of 150 to 456 isobaric PCs with different combinations of fatty acid chain lengths and degrees of desaturation for the two residues of which only a small subset of one to four PCs (fatty acid level) have been quantified by Lipidyzer$^{\text{TM}}$. Behind PCs of the fatty acid level, there are even more molecules on the level of chemical structures, due to differences in the order of both fatty acid chains, positions of double bounds, or stereo-chemistry. Although reliable interpretations require chemical structures, since similar compounds with little changes in respect to their configuration or desaturation lead to different biochemical functions, such as changes in their fluidity [136, 137], disturbances in the lipase activity [202], or go along with the mortality of patients suffering from cardiovascular diseases [135], PCs measured on the fatty acid level clearly reduce the number of possible constituents and therefore, enable more differentiating interpretations compared to PC sums of the lipid species level.

3.6.1 Strategies for structural elucidation of complex lipids

Although there are decision rule based (LipidMatch [188] and Lipid Data Analyzer (LDA) [189]) or comparative (Greazy [190] and MS-DIAL [191]) approaches that use analytic data to predict complex lipids, these methods cannot be applied to metabolites of MRM based FIA platforms, such as Absolute*IDQ*™, since these methods expect metabolites measured by non-targeted LC-MS devices. Moreover, these approaches assist characterization of complex lipids on the fatty acid level, but they are not capable to elucidate the main constituents of PC sums in biological tissues.

In contrast to these approaches, our method compares metabolite levels of samples that were measured on two platforms covering two different precision levels, such as the lipid species level and the fatty acid level. By this means, constituents of measured lipid sums are identified, and their quantitative contribution determined to serve imputation of other independent data sets.

An experimental (non-targeted) two-dimensional LC-MS technique, proposed by Holčapek et al. (2015) uses a reversed phase (RP) chromatography to separate lipid classes and a coupled HILIC to identify the lipid species [185]. Resulting fragment spectra facilitate identification complex lipids on the fatty acid level. Although they annotated 18 measured PCs with labels on the fatty acid level and on the lipid species level, their experimental design cannot resolve quantitative compositions. The focus was on the proof of the 2D-LC-MS measuring concept and on the study of the RT behavior of complex lipids. With our results, we can confirm that 14 of 18 measured PCs are among our specified main constituents. In the remaining four cases, a minor constituent was measured, or the measured PC was not available in our data set. In our study, we explained 38 measured PC compositions and moreover specified quantitative contributions of constituents.

3.6.2 Attributes of PC compositions in selected cohorts

Estimated PC compositions can be transferred to other samples and cohorts (by imputation of concentrations) under the assumption that factors f are relatively stable across samples. In this study, concentrations were measured in 56 plasma samples per subject that has been collected during specific timepoints and challenges (HuMet). Within this data basis we demonstrated that the factors f (i.e. distributions of quotients q) (containing 56 samples) of the main constituents are not significantly different between our four subjects for 18 out of 38 PC compositions. In samples of a second independent data set (QMDiab), we replicated factors f of the main constituents for 13 PC compositions and showed that distributions of quotients q between subsets containing male controls, female controls, and cases are not significantly different for 33 of 37 PC compositions. Therefore, only 4 PC compositions were significantly different between subsets. The study design of HuMet would even allow statistical tests regarding to the variation of factors f during challenges, indeed the small number of four subjects led to no significant results (the null hypothesis [H_0 : no difference] will not be rejected due to reduced statistical power). Moreover, statistical tests, including ANOVA or Kruskal-Wallis, determine empirical evidence about the difference of two or more distributions, but we were searching for equivalence. Thus, we were restricted to the phrase 'not significantly different' instead of 'significantly equal'. Because of only four samples per time point, we assessed the stability visually and concluded that there is a variation of factor f across challenges, but this variation is much smaller than the variation across subjects. Furthermore, estimated quantitative compositions are restricted to blood plasma. Ratios may be different in other matrices; however, plasma is the most frequently used tissue within metabolomics experiments [40]. In general, we demonstrated the reproducibility of PC compositions in our data; for an overarching statement further analyses in large heterogeneous data sets measured in various matrices by lipidomics platforms of the lipid species and fatty acid levels are necessary.

Besides small variation in factors f, we checked the consistency of variance that is explained by linear models and the values of f. In most PC compositions, large R^2 and large f are co-observed. However, these measures are contradicting in some PC compositions: e.g. PC aa C42:6 = PC 20:0_22:6 + R and PC aa C40:3 = PC 20:0_20:3 + R showed high factors f of 0.7 and 0.5 (indicating that measured constituents are main parts of both compositions), but very low R^2 of 5.9% and 0.2%, respectively. The increased variations in factors f of 0.85 and 0.49 (mean: 0.33, median: 0.11) suggest that this observation may be traced back to deviating measuring accuracies of both platforms or to individually flexible compositions of the respective PC sums.

Although most factors f are smaller than one, we found four unexpected clearly larger factors f (between 1.2 to 1.5). In these cases, measured concentrations of constituents on the fatty acid level are larger than the concentrations of the respective PC sums (lipid species level), which would not occur, if both platforms would measure the denoted compounds exactly. This deviation may be a consequence of measuring accuracy, external influences, probe handling, or different compound extraction procedures [78]. Although Absolute*IDQ*$^{\text{TM}}$ reports concentrations in absolute units, values are marked as semi-quantitative, because only a small number of internal standards is measured. We specified the 5% and 95% confidence interval for each factor f to indicate its variation but could not adjust factors f to a maximum of 1, because R that potentially contains several hundred non-measured compounds could not be determined on concentration level. However, Bland-Altman plots and linear models indicate that PCs measured on the fatty acid level that we classified as main constituents are important factors of respective PC sums (lipid species level) and that in most cases R is very small. Most of our PC sums consist of one main (measured) constituent measured on the fatty acid level.

3.6.3 Precision levels in lipidomics experiments

In general, assumptions regarding the composition of PCs of the lipid species level were derived from the abundance of cellular free fatty acids: e.g. the even numbered C16:0, C18:0, C18:1, C18:2 and C20:4 are common fatty acids in human plasma [138, 139, 203, 204]. Based on these assumptions, Biocrates$^{\text{TM}}$ provides a list containing probable $2-16$ PCs of the fatty acid level for PC sums measured on the lipid species level by Absolute*IDQ*$^{\text{TM}}$ kit [194]. PCs (fatty acid level) quantified by Lipidyzer$^{\text{TM}}$ cover a subset of one to four of these constituents per PC sum. Therefore, it is plausible that main constituents of the fatty acid level are among the measured compounds. With our results we can confirm that 47 of 55 PCs measured on the fatty acid level (mapped to 10 of 13 PC aa sums) contain at least one common fatty acid chain listed above. The remaining 8 PCs measured on the fatty acid level either contain C14:0 or C20:0.

However, we found notable exceptions: 8 PCs of the fatty acid level that were mapped to PC ae sums contain none of the mentioned common fatty acids. Most of these PCs consist of one odd-numbered fatty acid chain and C16, C20, or C22 with varying desaturation. Although Absolute*IDQ*$^{\text{TM}}$ MS technology cannot distinguish between the isobaric PC O-$x : y$ and PC $x - 1 : y$, we principally confirm that most PC ae sums do not consist of our measured isobaric PC (aa)s of the fatty acid level (mean $f = 0.16$). Interestingly, we identified three PC ae sums containing notable amounts of odd-numbered constituents (f between 0.20 and 0.34).

In the literature, biochemical interpretations of results containing PC sums (lipid species level) are regularly based on expected concrete underlying molecular structures [140, 141, 205]. As an example, Suhre et al. 2011 assumed that PC aa C36:4, PC aa C38:4, PC aa C38:5, and PC ae C38:3 are composed by one saturated or mono-unsaturated fatty acid C16 or C18 and of one poly-unsaturated fatty acid. For PC aa C30:0, PC aa C30:2, PC aa C32:0 and PC ae C34:1 they expect at least one C16 or C18 chain to be saturated or mono-unsaturated, but no poly-unsaturated fatty acids [140]. Now, we confirmed these expectations for 5 of 8 PC sums and even quantified related compositions: in mean these constituents contribute 49% (factor *f* between 0.23 and 1.02) to the concentrations of respective PC sums. Accordingly, these PCs (fatty acid level) describe in mean 62% (R^2 ranges from 35% to 83%) of the variance of respective PC sums (lipid species level). Relevant PCs of the fatty acid level for the remaining three PC sums have not been reported by Lipidyzer^{TM}. We could relativize the supposition of Floegel et al. 2013 that PC aa C32:1 mainly consists of PC 14:0_18:1 [141] and showed that this compound is among its minor constituents.

PC sums (lipid species level) and PCs of the fatty acid level belong to two different resolution levels within a range from the lipid class level to concrete chemical structures. PCs measured on the fatty acid level are still classes but not individual molecules, because the *sn*-1 and *sn*-2 positions are not distinguished and the positions of double bonds and their stereo-chemistry cannot be determined. However, specific side chains prefer one of the *sn*-1 or *sn*-2 positions of PCs in human plasma: in 95% of cases 16:0 is bound to the *sn*-1 position, 96% of 18:0 to position *sn*-1, 76% of 18:1 to position *sn*-2, 89% of 18:2 to position *sn*-2 and almost 100% of 20:4 to position *sn*-2 [139]. These variations have a large effect on the biochemical physiology in cells and tissues [135–137, 202]. Even if PCs annotated on the fatty acid level are not equivalent to concrete chemical structures, they already enable interpretations in a more precise biochemical context compared to PCs measured on the lipid species level with an unknown composition of chain lengths and double bonds or usage of assumptions regarding their composition. Our analysis enables biochemical interpretations on the fatty acid level for a large number of studies that led to data sets containing PCs measured on the lipid species level by Absolute*IDQ*^{TM}. On the one hand, determined quantitative compositions of PC sums can be used directly for imputation of measured PCs. On the other hand, based on a small subset of samples that are measured on both precision levels, our automated workflow can estimate factors *f* for imputation of metabolite concentrations of the remaining samples.

3.6.4 Summary

AbsoluteIDQ^{TM} is a well-established and efficient method for targeted quantification of lipids and amino acids. Recently, more and more precise MS based lipidomics platforms such as Lipidyzer$^{\text{TM}}$ appeared that quantify metabolites on a better resolution level. In this study, we specified mostly one single PC measured on the fatty acid level that is the main constituent of respective (24 of 38) PC sums (lipid species level) measured by AbsoluteIDQ^{TM} in human plasma and with metabolite levels of a second cohort, we replicated 13 and 19 quantitative compositions, respectively, by a direct comparison of factors f or a comparison of measured and imputed concentrations. Thereby, we enable more precise interpretation of scientific results incorporating PC sums, without the necessity of further MS analyses.

CHAPTER 4

Results and discussion II: Systems biology models for annotation of unknown metabolites

Metabolites with uncharacterized chemical identities measured by non-targeted metabolomics often mark dead ends of scientific analyses in metabolomics research. Unknown metabolites cover respective results and cannot be set into a biochemical context. To enable appropriate interpretations and further (e.g. clinical) usage of respective results, contained unknown metabolites need to be characterized.

I previously published major parts of Chapters 4.2 (data integration), 4.3 (automated prediction of pathways and reactions), and 4.6 (experimental verification of candidates) and minor parts of Chapter 4.5.4 (manual candidate selection) including respective materials, methods, and discussion; Quell et al. 2017 [150]. This work is significantly extended by contents of Chapters 4.4 (comparative annotation in different data sets and replication), 4.5 (automated candidate selection), and 4.7 (organization of (predicted) metabolite annotations).

4.1 Background

Although metabolite identification in wet-laboratories is especially successful for characterization of selected singular molecules, large sets of unknown metabolites require high-throughput *in silico* approaches (An overview of current popular methods is prepared in Table 4.1). The most basic approaches compare entire measured MS or MS/MS spectra of unknown metabolites to spectra deposited in public databases [83–85]. Since availability and comparability of fragmentation spectra depend on the measuring device, one of these approaches, MetAssign [86], clusters m/z and intensity of all peaks of measured and target spectra separately and ranks resulting

Method	**Comparative**	**FingerID & CSI:FingerID**	**SIMPLE**	**MetAssig**	**CFM**	**MetFrag**	**compMS^2 Miner**	**Meta-MapR**	**My approach**
References	Stein et al. 1994, Scheubert et al. 2013, Tautenhahn et al. 2012	Heinonen et al. 2012, Shen et al. 2014, Dührkop et al. 2015	Nguyen et al. 2018	Daly et al. 2014	Allen et al. 2015	Ruttkies et al. 2016	Edmands et al. 2017	Grapov et al. 2015	Quell et al. 2017 + extensions
	[83–85]	[90, 92]	[93]	[86]	[87]	[88]	[89]	[94]	[150]
Category									
searching spectral libraries	yes	—	—	yes	—	—	yes	—	—
in silico fragmentation	—	—	—	—	yes	yes	yes	—	—
machine learning	—	yes	yes	—	yes	—	—	—	—
searching structure DBs	—	yes	yes	—	yes	yes	—	—	yes
systems biology modeling	—	—	—	—	—	—	—	yes	yes
Integrates									
frag. spectra	yes	yes	yes	yes	yes	yes	yes	yes	—
DB reactions	—	—	—	—	—	—	—	yes	yes
DB relations	—	—	—	—	—	—	—	—	yes
RT	—	—	—	yes	—	yes	yes	—	yes
met. levels	—	—	—	—	—	—	—	yes	yes
chem. FP	—	yes	yes	—	—	—	—	yes	yes
DB mets.	yes	yes	yes	yes	yes	yes	yes	yes	yes
Predicts									
candidates	yes	yes	yes	yes	yes	yes	yes	—	yes*
bioch. PWs	—	—	—	—	—	—	—	—	yes
reactions	—	—	—	—	—	—	—	—	yes
chemical properties	—	yes	yes	—	—	—	—	—	—
biochemical context	—	—	—	—	—	—	—	yes	yes

Table 4.1: Properties of metabolite characterization approaches
Selected metabolite characterization approaches are based on spectral libraries, *in silico* fragmentation, machine learning, structure databases, or systems biology modeling. Most of the opposed approaches require fragmentation spectra, while reactions or metabolite-gene relations, metabolite levels and chemical fingerprints are only rarely used. The biochemical environment (e.g. biochemical pathways or reactions) are predicted by my approach only. *) prediction of specific candidate molecules is restricted to unknown metabolites with measured high-resolution mass.

candidates to improve comparability between spectra of different devices.

In order to become independent of spectral databases that are focused on specific measuring techniques or devices, approaches use *in silico* fragmentation of database structures or predict structural properties (e.g. molecular fingerprints) of unknown metabolites. As an example, CFM (competitive fragmentation modeling) [87] predicts peaks that are most likely observed for chemical structures by a Markov-process that subsequently estimates the likelihood of possible fragmentation events (fragmentation graph). As a result, this method predicts mass spectra given a chemical structure or ranks possible structures of public databases based on an MS/MS spectrum. MetFrag [88] applies *in silico* fragmentation and calculation of retention times (RTs) to substantiate candidates for unknown metabolites. The compMS^{2}Miner [89] combines several methods including *in silico* reconstruction of fragmentation, noise filtration, substructure annotation by fragmentation spectra and similarities to database spectra, functional group detection, and nearest neighbor chemical similarity scoring. A significantly different approach is pursued by FingerID [90], CSI:FingerID [91, 92], and SIMPLE [93], which predict molecular fingerprints (e.g. PubChem fingerprint) based on fragmentation spectra of unknown metabolites (in part supported by fragmentation tree computation) and query structure databases for compounds with similar fingerprints (similarity is estimated by the Tanimoto coefficient).

As a graph-based tool, the basic concept behind MetMapR [94] is closer to my approach than the other methods described above: Grapov et al. (2015) started with prior knowledge of biochemical reactions (KEGG RPAIR) to interconnect user defined known metabolites. To this network, edges between metabolites with similar PubChem substructure fingerprints are added. Finally, unknown metabolites are attached through similarities in pairwise mass spectra, or via empirical (parametric Pearson or non-parametric Spearman) correlations determined in measured metabolite levels. Although Grapov et al. construct biochemical networks incorporating unknowns, automated annotation of metabolites is not included.

Overview of the metabolite characterization approach introduced

In this work, I propose a systems biology approach for characterization of unknown metabolites: By estimation of pairwise partial correlations (Gaussian graphical model (GGM)) between measured metabolite levels, unknown metabolites are directly incorporated during creation of the network models' backbone reconstructing biochemical pathways. Subsequently, I further extend the network by metabolite-gene associations (mGWAS) and prior biochemical knowledge from public databases. In Addition, I provide modules accessing the constructed network model to transfer annotations from known to unknown metabolites, predict pathways, select reactions that connect known and unknown metabolites, and choose specific candidate molecules (for unknown metabolites with measured high-resolution mass) based on database structures that can then be tested experimentally.

I applied my approach to unknown metabolites in metabolomics data measured in blood of 2 279 subjects of SUMMIT [19–23] by the established non-targeted MS platform of Metabolon$^{\text{TM}}$ that has been widely used for metabolite quantification in large cohorts. To demonstrate the automated candidate selection module, I used unknown metabolites measured in 1 209 samples of the GCKD cohort by the new non-targeted high-resolution platform of Metabolon$^{\text{TM}}$. As a proof of principle, I compared automatically generated annotations of 26 unknown metabolites captured by both platforms, determined benefits of automated annotations by combing network models of both data sets, and tested selected candidates on the LC-MS metabolomics platform for verification. I finally propose a database structure for organization of known, predicted, or verified information among unknown metabolites to store annotations determined in various data sets, to find consistent or contradictory annotations, and to access annotations for sets of unknown metabolites of future metabolomics studies.

A running open source toolbox of this approach including required input files, sample output files, and instructions is provided in Supplemental File B.1.

4.2 Integration of data from different biological domains

The network model is the core part of the approach and the basis of analytic and predictive methods (Figure 4.1). It connects unknown metabolites to complementary functional information from heterogeneous data resources (including data driven GGMs, associative mGWAS, or data base relations) and allows automated mining of these connections for metabolite identification. As a consequence, edges in the network represent various types of relations, which are integrated into the network in separate steps using data type specific thresholds (Figure B.2). 637 (388 known, 249 unknown) of the 758 metabolites measured by the established Metabolon$^{\text{TM}}$ platform in 2 279 blood samples of the SUMMIT cohort are connected by 1 040 GGM edges leading to a network model with one large connected component and 17 separate sub graphs with a maximum of 7 vertices. Adding genetic associations from

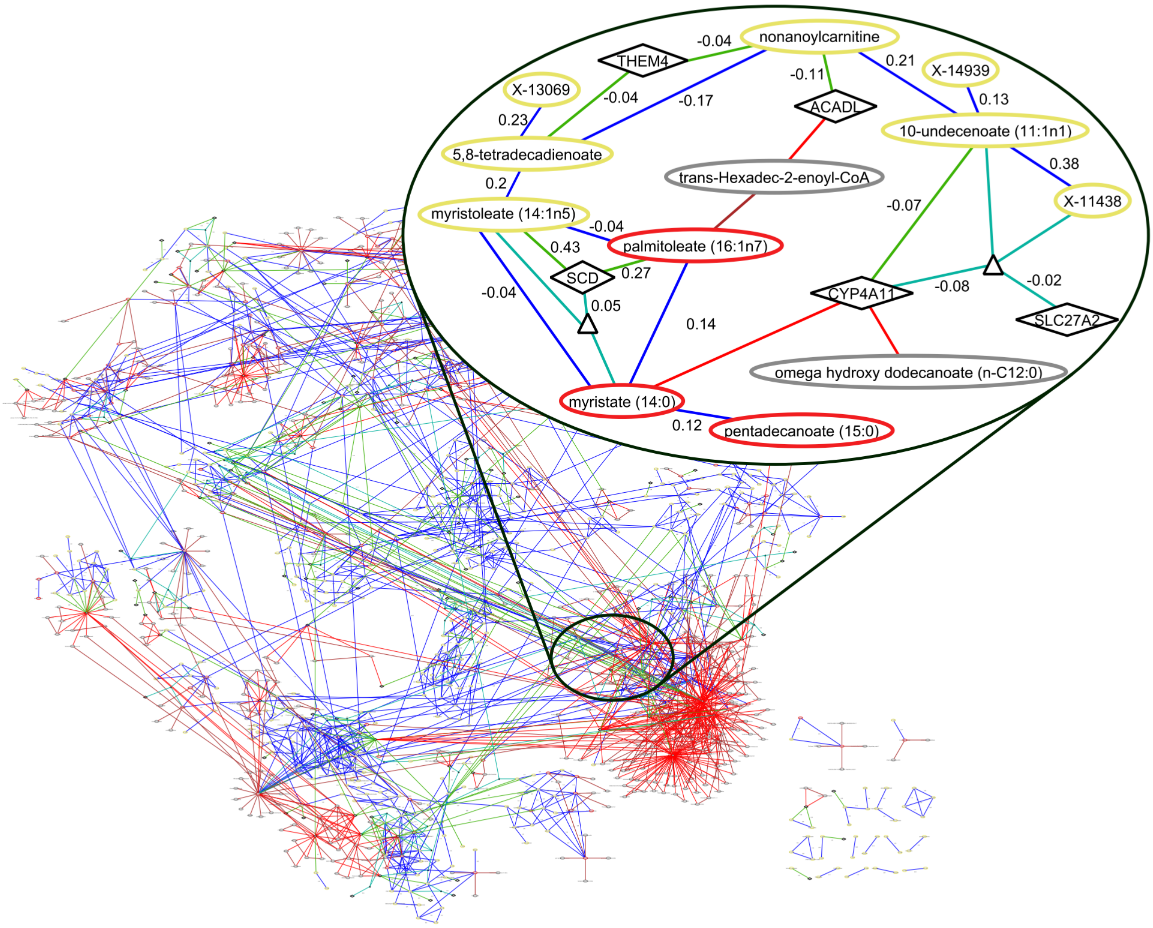


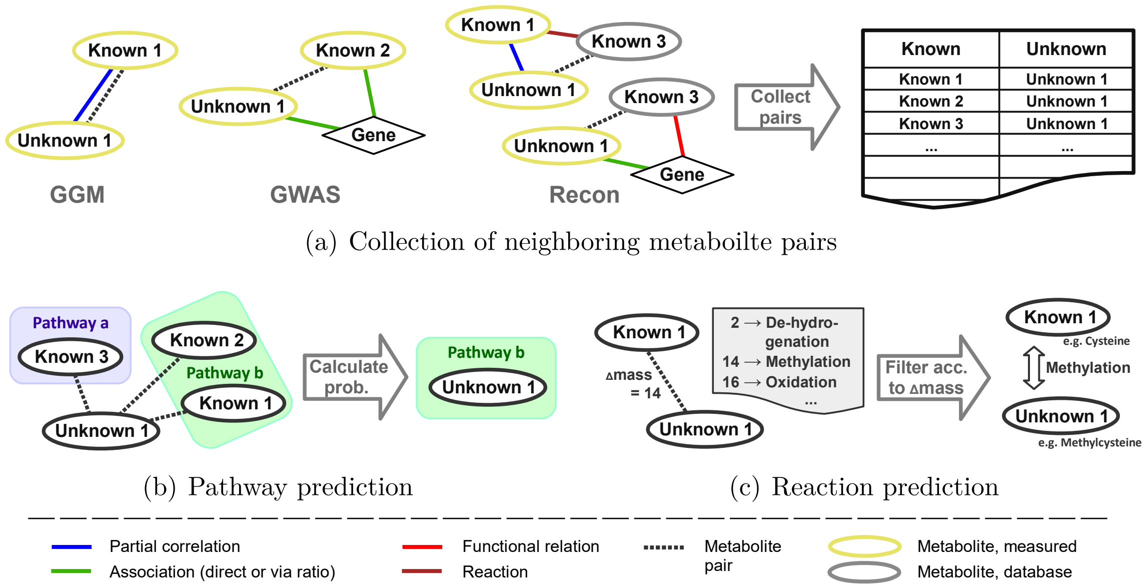
Figure 4.1: Graphical representation of the integrated network model
The final network model embeds 254 measured unknown metabolites into their biochemical and functional context. Edge colors indicate the type of connection as follows: GGM (blue), mGWAS (green), functional relation by Recon (red), reaction by Recon (brown). GGM edges are labeled by the mass difference of metabolites and the β is shown for mGWAS edges. The shape of nodes indicates the element type: metabolite (oval), gene (diamond) and the border color of nodes indicates if it is a measured metabolite (yellow), a Recon metabolite (grey), or both (red). Non-connected metabolites and genes are not shown. I previously published a similar version of this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

published mGWAS to the network, 186 measured metabolites (136 known, 50 unknown) are linked to 134 genes (169 metabolites directly, 73 metabolites through ratios, and 56 through both). 175 metabolites are connected through a GGM edge as well as through mGWAS edges via a gene, thus 648 measured metabolites (394 known, 254 unknown) of our network model are connected. We then integrated 480 metabolites and mapped 139 metabolites and 57 genes of the public database Recon (Virtual Metabolic Human Database). 591 of those metabolites are connected through 1 152 reactions, of which 343 are functionally related to 57 genes. Only a subset of 139 of our measured metabolites in the network could be directly mapped to metabolites in Recon. In the final model, 181 unknown metabolites are connected by a GGM edge (171) or via a gene (35) to known metabolites. A graphml file (XML based file format for storing graphs; visual representation by publicly available software such as yEd) of the network model is prepared in Supplemental File B.2.

The overall network, which shows a scale-free topology, embeds unknown metabolites into their biochemical and functional context (Figure 4.1, zoom in) by connecting them to known metabolites via direct edges (GGM, blue) or indirect edges (mGWAS, green; Recon, red). Thereby metabolites are connected neglecting the compartmentalization (organ, tissue) of biochemical processes. In contrast to networks aiming at a realistic reconstruction of complete human metabolism for modeling and simulation, the network generated in our study is supposed to capture as many functional links between metabolites as possible to provide hints for metabolite identification irrespective of their exchange between compartments or organs.

4.3 Automated workflow-based characterization of unknown metabolites

The neighborhood of unknown metabolites (consisting of known, annotated metabolites) is a common input source for predictions of pathways and reactions. For this reason, we defined neighborhood of metabolites in the network model (Figure 4.2(a)) and demonstrated these modules based on metabolites of SUMMIT that were measured on the established Metabolon^{TM} platform: 171 and 35 unknown metabolites are neighbors of known metabolites through GGM edges and through common associated genes (mGWAS), respectively; 68 unknown metabolites are neighbors of known metabolites, because of an associated or related third metabolite or gene. In total, 182 unknown metabolites share 412 neighboring known metabolites. Our modules for automated prediction of pathways and reactions prefer data sets with a balanced number of known and unknown metabolites.



(a) Collection of neighboring metaboilte pairs

(b) Pathway prediction

(c) Reaction prediction

Figure 4.2: Neighborhood definition and automated annotation schema
For the prediction of pathways and reactions of unknown metabolites, (a) neighbors of each metabolite were collected based on direct partial correlation (GGM) edges, common genetic links by mGWAS, or functional connections via a Recon metabolite or gene. The statistics (b), which is calculated among the neighborhood of known metabolites, is applied to each unknown metabolite to predict the most probable pathway. For the reaction prediction, (c) reactions are assigned to pairs of neighboring metabolites based on their mass difference Δm and a list of typical mass differences that are characteristic for specific reactions. While node labels in the figure indicate unknown or known metabolites known-known neighbors are used for validation of the prediction approach, and unknown-unknown pairs are also analyzed in the reaction prediction. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

4.3.1 Prediction of pathways

Based on known metabolites in the neighborhood of unknown metabolites, we were able to automatically predict super pathways for 180 and sub pathways for 178 out of 183 unknown metabolites that were connected to at least one known metabolite either directly or indirectly via a gene or a third metabolite (Figure 4.2(b)). To 150 metabolites a clear sub pathway was assigned. For further 28 metabolites two (mostly similar) sub pathways were suggested. Table 4.2 summarizes the proportion of predicted super and sub pathways per confidence class. In general, unknown metabolites that are not well connected or belong to a cluster of other unknown metabolites can be predicted with only low confidence. From a methods view, the confidence class only counts for the super pathway prediction, but as shown in Table B.1 the sub pathway prediction behaves similarly. Consequently, the confidence of predicted sub pathways increases with a rising confidence of predicted super pathways. A complete list of predicted super and sub pathways for unknown metabolites is provided in Table B.2 and the generic output of the annotation workflow is provided in Supplemental File B.3. Our approach is able to classify pathways for large amounts of unknown metabolites automatically.

Confidence [%]	Confidence label	Prediction rate [%] super pathway	Count super pathway	Count (>1 option) sub pathway
≥ 97.5	very high	18.3	33	33 (6)
≥ 95	high	2.2	4	4 (0)
≥ 90	medium	35.0	63	62 (16)
≥ 85	low	31.1	56	55 (5)
< 85	very low	13.3	24	24 (1)

Table 4.2: Proportion of unknown metabolites with predicted pathways
Pathway prediction on the five confidence levels were determined based on thresholds among known metabolites (see Materials and Methods). I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

4.3.2 Prediction of reactions

Prediction of reactions, such as methylation, oxidation, hydroxylation, phosphorylation, carboxylation, or hydrogenation (Table 4.3), between known and unknown metabolites enables the *in silico* application of reactions to known metabolites to select specific candidates for unknown metabolites.

We tried the simple approach of assigning reactions to pairs of neighboring metabolites, according to relations in the network model, which we assumed to be connected by a reaction (Figure 4.2(c)) based on a typical, reaction-specific change in mass. Thereby, we considered all pairs with mass difference Δm_{pair} within an error interval of $\Delta m_{expected} \pm 0.3$ to compensate for the limited mass resolution in our metabolomics data set. As Breitling et al. [148] showed that the accuracy of reaction prediction significantly depends on the mass resolution of measuring devices, this interval should be adjusted for platforms with better resolution to allow an improved differentiation of reactions. Table 4.3 shows the number of pairs of known, known/unknown, and unknown metabolites per assigned reaction. In total, we assigned 100 reactions to pairs of unknown and known metabolites. We predicted reactions also between 45 pairs

Reaction	Δmass [Da]	known	known	unknown
		known	unknown	unknown
(Oxidative) deamination	1	31 (23)	6	3
(De)hydrogenation	2	79 (53)	15	11
(De)methylation or Alkyl-chain-elongation	14	63 (37)	8	9
Oxidation, Hydroxylation, or Epoxidation	16	75 (48)	14	4
(De)ethylation or Alkyl-chain-elongation	28	64 (37)	5	7
Quinone, CH3 to COOH, or Nitro reduction	30	24 (5)	4	1
Bis-oxidation	32	24 (3)	9	0
(De)acetylation	42	29 (5)	13	2
(De)carboxylation	44	26 (10)	13	1
Sulfation or Phosphatation	80 or 96	23 (9)	5	4
Taurine Conjugation	107	2 (0)	1	1
Cys Conjugation	121 or 119	3 (1)	3	1
Glucuronidation	176 or 192	10 (4)	3	1
GSH Conjugation	307 or 305	0 (0)	1	0
$\sum$ assigned reactions		453 (235)	100	45
$\sum$ neighbors in the network model		5 600	899	223

Table 4.3: Summary of assigned reactions based on Δmass and neighborship
Selected frequently occurring reactions were assigned to pairs of metabolites (known-known, known-unknown, or unknown-unknown) according to their typical change in mass ($\Delta m \pm 0.3$). The neighborhood definition leads to collection of pairs of metabolites within the network model (connected via GGM, mGWAS, or Recon edges). The number in brackets in the column of reactions between pairs of neighboring known metabolites indicates the number of verified reactions. I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

of unknown metabolites as it can serve as one element of metabolite characterization. A complete list of assigned reactions with unknown metabolite is prepared in Table B.3 and Supplemental File B.3 provides the untreated the output of the annotation workflow. During a verification among known metabolites, the assignment procedure leads to $58-74\%$ correct predictions for reactions with Δmass lower than 30. With 23 true out of 31 assigned (de)amination processes (74%) and 53 true out of 79 assigned (de)hydrogenation processes (67%), the simplified reaction prediction approach worked best for these two types of reactions in our data (Table 4.3).

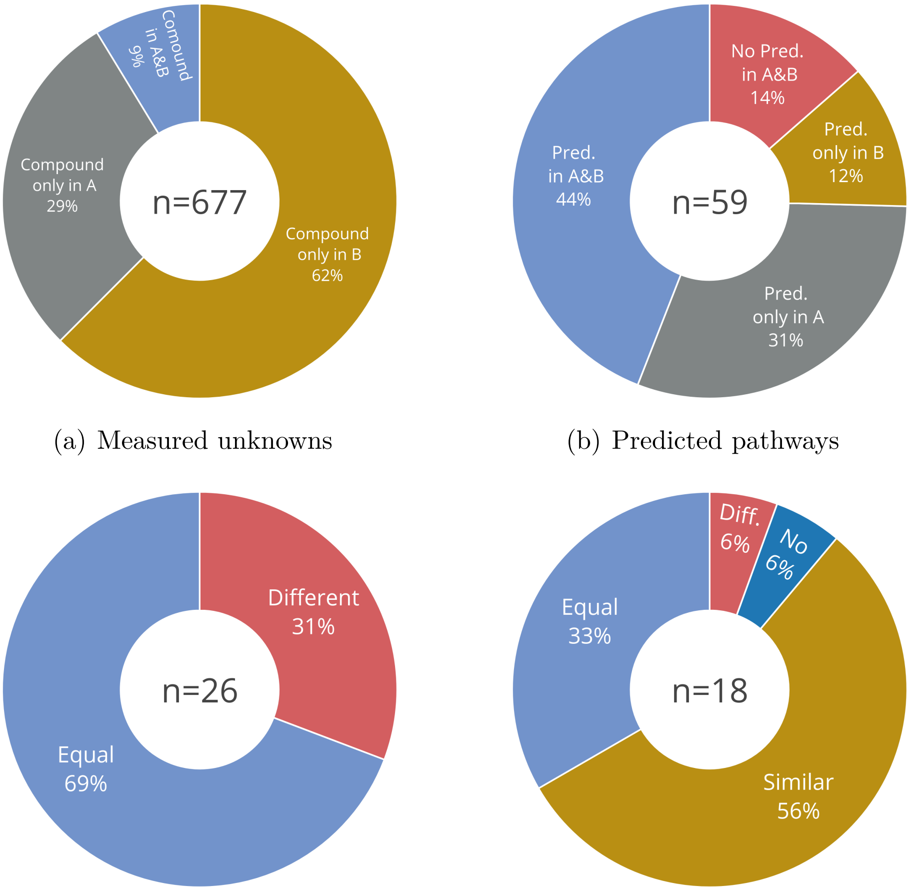
Additionally, we applied a second prediction criterion considering the retention index (RI) of reactants for the most frequently predicted reaction type, the dehydrogenation reaction. To this end, we used the distribution of differences in RI between pairs of known metabolites that are connected by a dehydrogenation reaction compared to known metabolites with the same Δmass but connected via another reaction (Figure B.3). Out of 15 pairs of connected unknown and known metabolites with $\Delta \mathrm{m} = 2 \pm 0.3$ and $\Delta \mathrm{RI} \leq 355.9$, we classified 12 pairs to be part of a dehydrogenation based on this additional criterion (Table B.4). Beyond that, we also considered 11 pairs of unknown metabolites to learn about the relationship among themselves, of which 7 pairs were predicted to be connected by a dehydrogenation.

4.4 Comparative annotation of metabolites measured on different platforms

We had access to a second independent data set of 1 382 metabolite levels (750 known, 632 unknown) quantified in 1 209 blood samples of the GCKD cohort by the new non-targeted high-resolution LC-MS platform of Metabolon$^{\mathrm{TM}}$, facilitating comparisons of independent annotations for unknown metabolites and automatic selection of candidate molecules.

986 of 1 382 metabolites of the GCKD data set with up to 20% missing values went through the imputation step that is required for estimation of pairwise partial correlations (GGM estimation), and were forwarded to the automatic integration and annotation modules described in Chapters 4.2 and 4.3. The resulting network model contains 1 032 measured metabolites (550 known, 482 unknown) (note that additional 46 metabolites, not used for GGM generation, were included by the mGWAS), of which 879 metabolites (447 known, 432 unknown) are connected by 1 083 data driven GGM edges and 396 metabolites (217 known, 179 unknown) are associated to 435 genes (by 704 metabolite-gene associations; mGWAS). Integration of information from the public database Recon led to 689 metabolites (196 mapped to measured metabolites) connected through 665 reactions, and 376 metabolites (including 51 measured metabolites) functionally related to 56 genes (609 metabolite-gene relations). A subset of 208 measured metabolites could be mapped to compounds of Recon. In the final network model 1 789 pairs of 282 unknown metabolites are related to 647 neighboring known metabolites: 206 unknowns are connected directly via a GGM edge, 86 unknowns via an associated gene (mGWAS), and 97 unknowns via a functionally related and associated third metabolite or gene. A subset of 59 unknown metabolites was measured on both platforms (in the SUMMIT and GCKD

cohorts) and mapped by Metabolon$^{\text{TM}}$ (Figure 4.3(b)). A visual representation of the



(a) Measured unknowns

(b) Predicted pathways

(c) Predicted super pathways

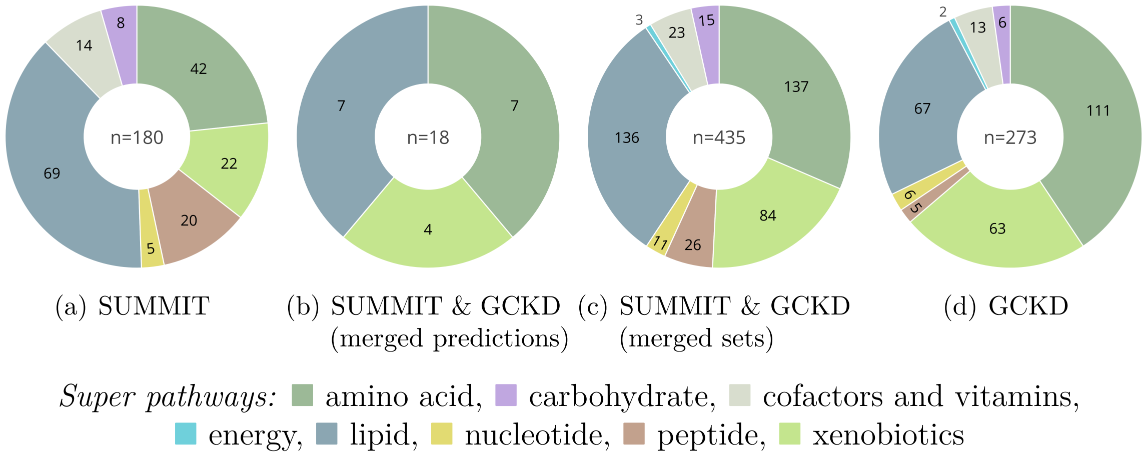
(d) Predicted sub pathways

Figure 4.3: Quantities of predicted pathways in SUMMIT and GCKD
(a) 59 of 677 mapped unknown molecules (9%) were measured by both Metabolon$^{\text{TM}}$ platforms in SUMMIT (A) and GCKD (B). (b) For 26 of those 59 commonly measured unknown metabolites (44%) super pathways were independently predicted in both sets. (c) 18 of 26 super pathways (69%) were predicted equally and (d) 6 of 18 sub pathways (with matching super pathways) were predicted equally and further 10 pathways were predicted similarly in both sets.

resulting network model is prepared in Supplemental File B.4. 1 682 of 1 965 nodes are connected in one large coherent scale-free structure. The remaining 283 nodes belong to 77 components of 2 to 17 nodes without links to the main network.

4.4.1 Comparison of separately performed annotations

Our automated approach annotates 273 and 270 unknown metabolites with super (Table B.5) and sub pathways (Table B.6), respectively (Figure 4.4(d)). According to



(a) SUMMIT

(b) SUMMIT & GCKD (merged predictions)

(c) SUMMIT & GCKD (merged sets)

(d) GCKD

Figure 4.4: Composition of predicted super pathways in SUMMIT and GCKD
The composition of predicted super pathways varies in both sets: (a) SUMMIT and (d) GCKD. (b) 18 of 26 super pathways were predicted equally in both sets consists of amino acids, lipids, and xenobiotics. (c) If the data sets were merged before data integration, pathways for 16 additional unknown metabolites were predicted.

Δmass (± 0.05) of neighboring metabolites (cf. Figure 4.2(a)), reactions were selected for 297 unknown metabolites to 210 known and 352 unknown metabolites. The second filtering criterion (based on ΔRI) was passed by 56 of 90 cases leading to the prediction of dehydrogenation reactions. Described annotations are summarized in the generic workflow output in Supplemental File B.5.

We compared annotations of 59 of 677 unknown metabolites that were measured by both platforms in SUMMIT and GCKD (Figure 4.3(a & b) and Table B.5). A subset of 26 of these unknown metabolites carry predicted super pathways in both data sets (Table 4.4), of which 18 annotations (69%) match (Figure 4.3(c)). Matching super pathways consist of 7 amino acids, 7 lipids, and 4 xenobiotics (Figure 4.4(b)). Noting, that in both sets the a priori probabilities of super pathways among known metabolites varies clearly. While most frequent super pathways of SUMMIT include lipids (38%), amino acids (23%), xenobiotics (12%), and peptides (11%), metabolites of GCKD are primarily classified as amino acid (41%), lipid (25%), or xenobiotics (23%) (Figure 4.4(a & d)). If equal and similar (cf. definition in Chapter 2.2.7) annotations are considered jointly, the numbers of predicted super pathways correspond to those of 16 of 24 matching (67%) predicted sub pathways (Figure 4.3(d)).

Metabolite	SUMMIT	GCKD	Equal
X-12093	amino acid	amino acid	e / e
X-12511	amino acid	amino acid	e / e
X-13835	amino acid	amino acid	e / e
X-11334	amino acid	amino acid	e / s
X-11787	amino acid	amino acid	e / s
X-15461	amino acid	amino acid	e / s
X-12704	amino acid	amino acid	e / –
X-16071	amino acid	lipid	d
X-12543	amino acid	xenobiotics	d
X-14095	amino acid	xenobiotics	d
X-17685	amino acid	xenobiotics	d
X-14056	cofactors and vitamins	amino acid	d
X-11440	lipid	lipid	e / s
X-11470	lipid	lipid	e / s
X-11540	lipid	lipid	e / s
X-13866	lipid	lipid	e / s
X-14662	lipid	lipid	e / s
X-16654	lipid	lipid	e / s
X-14626	lipid	lipid	e / d
X-02249	lipid	amino acid	d
X-15492	lipid	energy	d
X-12730	xenobiotics	xenobiotics	e / e
X-12847	xenobiotics	xenobiotics	e / e
X-13728	xenobiotics	xenobiotics	e / e
X-15728	xenobiotics	xenobiotics	e / s
X-15497	xenobiotics	lipid	d

Table 4.4: Super pathways predicted commonly in SUMMIT and GCKD
For 26 unknown metabolites, super and sub pathways were independently predicted based on both cohorts: SUMMIT and GCKD. The last column indicates, if super/sub pathways were predicted equally (e), differently (d), similarly (s), or not predicted (–). Entries are ordered by predicted super pathways and evaluation of sub pathways.

4.4.2 Benefit of annotations within merged sets

Since a subset of 227 metabolites (168 known, 59 unknown) was measured in SUMMIT as well as in GCKD, we analyzed if automatically generated annotations profit from merging two data sets of different Metabolon^{TM} platforms. To this end, we united the metabolite sets of SUMMIT, GCKD, and Recon and integrated data driven GGMs successively. Note that partial correlations of both sets were estimated separately to prevent biases resulting from different non-targeted measuring platforms. Data integration and annotation of unknown metabolites was performed by the workflow as described in Chapter 4.2.

The resulting network model (graph in Supplemental File B.6) contains 1 431 measured metabolites (754 known, 677 unknown): 2 043 GGM edges connect 1 314 metabolites (686 known, 628 unknown) and a subset of 547 metabolites (324 known, 223 unknown) is associated to 529 genes (978 mGWAS edges). In addition, 831 database-metabolites (222 measured) are connected by 847 reactions of Recon. Finally, 484 database-metabolites (71 measured) are functionally related to 95 genes (842 metabolite-gene relations). In the resulting network model that consists of one large coherent structure (2 420 nodes) and 66 encapsulated components of 2 to 8 nodes (200 nodes in total) with no connection to the main network, 443 unknown metabolites are connected to 887 known metabolites by 2 811 neighborship relations: 359 unknowns via GGM edges, 130 unknowns through a common associated gene (mGWAS), and 173 unknowns over a common gene or third metabolite of Recon that is associated or functionally related.

The automatic annotation modules led to the prediction of 435 super and 434 sub pathways for unknown metabolites (Tables B.5, and B.6). In addition, 782 reactions were selected (based on Δmass ± 0.3 according to the mass resolution of the older platform) for 384 unknown metabolites, of which 76 of 121 dehydrogenation reactions passed the more specific ΔRI criterion. Raw annotations are prepared Supplemental File B.7.

We compared pathway annotations of three sets: SUMMIT/GCKD (merged set), SUMMIT, and GCKD. Our special interest was on 51 of 59 unknown metabolites that carry pathway annotations and were mapped in both sets. For 427 of 677 unknown metabolites, pathways were predicted in at least two of the three sets (Figure 4.5(a) and Table B.5). Only 26 unknown metabolites had super pathway predictions in all three sets, of which 17 predictions were matching: 6 amino acid, 7 lipids, and 4 xenobiotics (Figure 4.5(a & b) and Table 4.5). The remaining 9 predictions are equal in two of the three sets. Matching pathways are primarily among the most frequent pathways of both cohorts: amino acids, lipids, xenobiotics (Figure 4.4(a & d)). Respective predictions of sub pathways match in 6 and 18 cases in all three or in two sets, respectively (Table B.7). The remaining two cases had no sub pathway prediction in one set. In total, 407 (95%) super and 377 (88%) of 427 sub pathways that were predicted in two of the three sets match (Figure 4.5(b & c) and Tables B.5 and B.6). Further 29 predicted sub pathways are similar, in two sets.

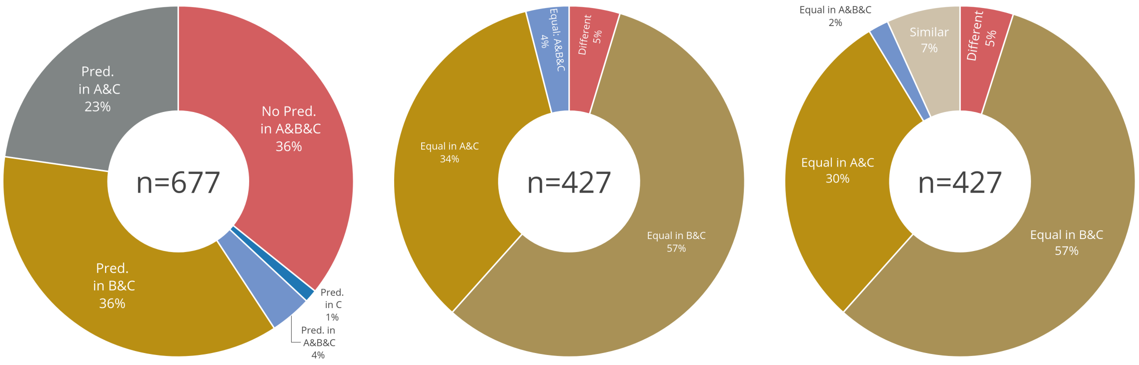
Metabolite	SUMMIT/GCKD	SUMMIT	GCKD	Equal
X-11334	amino acid	amino acid	amino acid	e / e / e
X-11787	amino acid	amino acid	amino acid	e / e / e
X-12093	amino acid	amino acid	amino acid	e / e / e
X-12511	amino acid	amino acid	amino acid	e / e / e
X-13835	amino acid	amino acid	amino acid	e / e / e
X-15461	amino acid	amino acid	amino acid	e / e / e
X-14314	amino acid	—	amino acid	– / – / e
X-14095	amino acid	amino acid	xenobiotics	d / e / d
X-14056	amino acid	cofactors and vitamins	amino acid	d / d / e
X-01911	amino acid	cofactors and vitamins	—	– / d / –
X-14302	amino acid	peptide	—	– / d / –
X-12007	carbohydrate	carbohydrate	—	– / e / –
X-12206	cofactors and vitamins	cofactors and vitamins	—	– / e / –
X-17162	cofactors and vitamins	cofactors and vitamins	—	– / e / –
X-15492	energy	lipid	energy	d / d / e
X-11440	lipid	lipid	lipid	e / e / e
X-11470	lipid	lipid	lipid	e / e / e
X-11540	lipid	lipid	lipid	e / e / e
X-13866	lipid	lipid	lipid	e / e / e
X-14626	lipid	lipid	lipid	e / e / e
X-14662	lipid	lipid	lipid	e / e / e
X-16654	lipid	lipid	lipid	e / e / e
X-16947	lipid	lipid	—	– / e / –
X-17299	lipid	lipid	—	– / e / –
X-17438	lipid	lipid	—	– / e / –
X-12822	lipid	—	lipid	– / – / e
X-12846	lipid	—	lipid	– / – / e
X-12860	lipid	—	lipid	– / – / e
X-15486	lipid	—	lipid	– / – / e
X-02249	lipid	lipid	amino acid	d / e / d
X-16071	lipid	amino acid	lipid	d / d / e
X-15497	lipid	xenobiotics	lipid	d / d / e
X-11429	nucleotide	nucleotide	—	– / e / –
X-12844	nucleotide	nucleotide	—	– / e / –
X-12216	peptide	amino acid	—	– / d / –
X-12730	xenobiotics	xenobiotics	xenobiotics	e / e / e
X-12847	xenobiotics	xenobiotics	xenobiotics	e / e / e
X-13728	xenobiotics	xenobiotics	xenobiotics	e / e / e
X-15728	xenobiotics	xenobiotics	xenobiotics	e / e / e
X-11452	xenobiotics	xenobiotics	—	– / e / –
X-12230	xenobiotics	xenobiotics	—	– / e / –
X-12231	xenobiotics	xenobiotics	—	– / e / –
X-12329	xenobiotics	xenobiotics	—	– / e / –
X-12407	xenobiotics	xenobiotics	—	– / e / –
X-12830	xenobiotics	xenobiotics	—	– / e / –
X-17185	xenobiotics	xenobiotics	—	– / e / –

Table continued (p. 2/2)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD	Equal
X-09789	xenobiotics	—	xenobiotics	– / – / e
X-12212	xenobiotics	—	xenobiotics	– / – / e
X-12543	xenobiotics	amino acid	xenobiotics	d / d / e
X-17685	xenobiotics	amino acid	xenobiotics	d / d / e
X-12704	xenobiotics	amino acid	amino acid	d / d / d

Table 4.5: Super pathways predicted in merged sets of SUMMIT and GCKD
Super pathways were predicted for 51 of 59 unknown metabolites that are measured in SUMMIT and in GCKD. The last column indicates, if super pathways were predicted equally (e) or differently (d): in SUMMIT/GCKD, SUMMIT, and GCKD; in SUMMIT/GCKD and SUMMIT; in SUMMIT/GCKD and GCKD; or if there is not prediction (–). Respective sub pathway predictions are prepared in Table B.7. Entries are ordered by super pathways predicted in SUMMIT/GCKD and by number of equally predicted pathways.

Interestingly, we found that super and sub pathways to additional 8 unknown metabolites (2 unknowns of SUMMIT and 6 unknowns of GCKD) were only predicted in the merged set, SUMMIT/GCKD (Figure 4.5(a)): 5 lipids, 1 energy, 1 peptide, and 1 xenobiotics (Table 4.6). Merging both data sets enables annotations for those unknown metabolites, if: (*i*) an unknown metabolite is measured in data set A only; (*ii*) but is not connected to a known metabolite in data set A; (*iii*) however, the



(a) Predicted pathways (b) Equally predicted super pw (c) Equally predicted sub pw

Figure 4.5: Quantities of commonly predicted pathways in SUMMIT and GCKD
(a) Proportion of unknown metabolites with predictions of super pathways in SUMMIT (A), GCKD (B) and SUMMIT/GCKD (C) sets. 26 unknown metabolites (4%) have predicted pathways in all sets, A, B, and C, and further 8 unknown metabolites (1.2%) have predictions only in the merged set C. (b and c) 427 of 677 unknown metabolites carry predicted pathways in at least two sets. In A and C, or in B and C large fractions of predicted super and sub pathways are equal.

Metabolite	Predicted super pathway	Predicted sub pathway	Measured in
X-06227	energy	oxidative phosphorylation	SUMMIT
X-13431	lipid	carnitine metabolism	GCKD
X-21846	lipid	carnitine metabolism	GCKD
X-24798	lipid	carnitine metabolism	GCKD
X-17705	lipid	sterol/steroid	GCKD
X-24348	lipid	sterol/steroid	GCKD
X-24834	peptide	fibrinogen cleavage peptide	GCKD
X-12092	xenobiotics	chemical	SUMMIT

Table 4.6: Pathways predicted in only in the merged set of SUMMIT and GCKD Super and sub pathways of 8 unknown metabolites were only predicted in the merged data set SUMMIT/GCKD, but not within the individual sets. Entries are sorted by the predicted super and sub pathways.

unknown shares a commonly associated gene with a known metabolite of data set B; (*iv*) which leads to neighboring unknown and known metabolites across both sets, A and B: e.g. X-24834 is only measured in the GCKD set and is associated to two genes, 'FUT2' and 'ABO', located in a subnetwork of three nodes with no connection to the main network (Figure B.4(a)). In the network model of SUMMIT, these genes 'FUT2' and 'ABO' are associated to 4 di- or poly-peptides (directly or via metabolite ratios): leucylalanine, ADpSGEGDFXAEGGGVR, ADSGEGDFXAEGGGVR, and DSGEGDFXAEGGGVR (Figure B.4(b)). Consequently, the unknown metabolite, X-24834, is neighbor of these known metabolites in the merged set, SUMMIT/GCKD, and annotated as peptide (Figure B.4(c) and Table 4.6).

4.5 Automated selection of candidate molecules

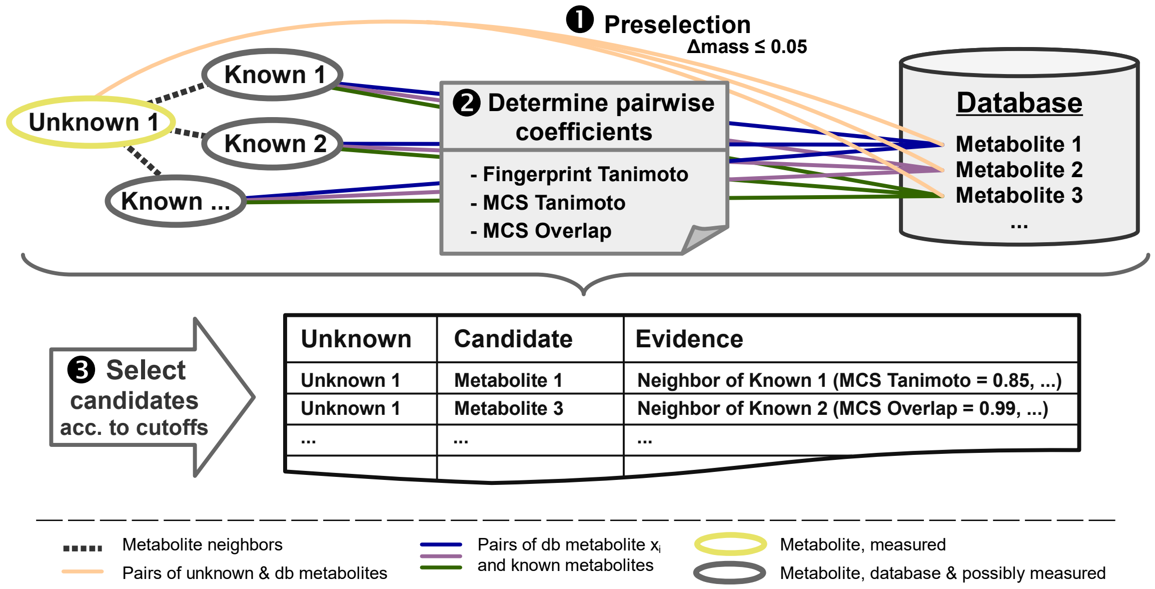
The availability of 1 382 metabolite levels (750 known, 632 unknown) measured by a high-resolution LC-MS platform of Metabolon^{TM} within 1 209 samples of the GCKD cohort enables more specific annotations and allows the automatic selection of possible candidate molecules among a pool of database-structures. We used molecules of Recon and HMDB to generate a suitable pool of structures.

As a pilot study for the following automated selection of candidate molecules, we searched HMDB [71] by the measured mass (m/z ± 0.01 (high-resolution platform), or ± 0.3 (established platform)) and examined the plausibility of the structure in the biochemical context of the network model. This procedure yields selection of 36 specific candidate molecules for 33 unknown metabolites based on HMDB (Table B.8).

One of these candidates, dimethyl-lysine for X-24983, was confirmed in the meantime by Metabolon^{TM} for proof of concept.

4.5.1 Automated procedure

To receive a large number of neighboring known metabolites for the set of high-resolution unknowns, we used the network model of Chapter 4.4.2 that integrates data driven information from SUMMIT and GCKD, published associations of KORA F4, and functional relations of Recon. The final network model contains 1 431 metabolites (754 known, 677 unknown) and 529 genes. 443 of 677 unknown metabolites share 756 neighboring known metabolites (according to neighborhood definition, Figure 4.2(a)). The subset of 315 unknown metabolites that were measured on the high-resolution device and were connected to known metabolites in the network model went through our three steps of the automated candidate selection procedure: pre-selection according to Δmass, estimation of structural similarity, and filtering of candidates passing a (similarity-) cutoff (Figure 4.6).



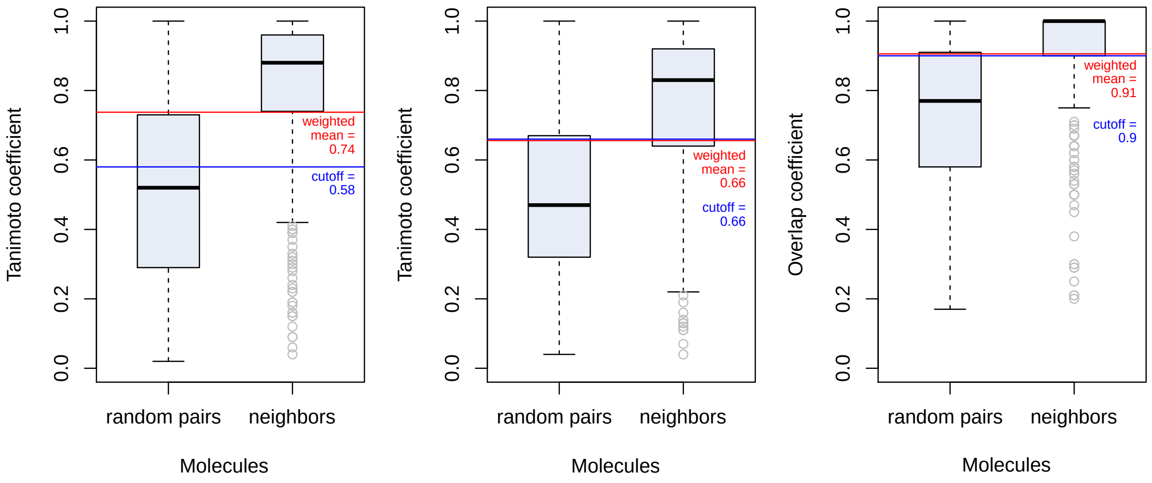
Unknown	Candidate	Evidence
Unknown 1	Metabolite 1	Neighbor of Known 1 (MCS Tanimoto = 0.85, ...)
Unknown 1	Metabolite 3	Neighbor of Known 2 (MCS Overlap = 0.99, ...)
...	...	...

Figure 4.6: Schema of automated selection of candidate molecules
The preselection of potential candidate molecules is performed by a database search (Δmass $\leq$ 0.05) using the high-resolution mass of unknown molecules. In a second step the fingerprint Tanimoto, MCS Tanimoto and MCS Overlap coefficients of the chemical structures of pairwise neighboring known and potential candidate metabolites were estimated. Resulting candidate molecules were selected according to empirical determined cutoffs.

The approach follows the observation that neighboring metabolites of the network model are structurally more similar (fingerprint or maximum common substructure (MCS)) than random pairs (Figure 4.7). After searching public metabolic databases, such as HMDB or Recon by the measured mass of unknown metabolites ($\Delta\text{mass} \leq 0.05$) to pre-select possible database candidates, the chemical structures of pairs of pre-selected candidate molecules and known metabolites that are neighbors of the respective unknown metabolite are compared by three similarity measures: fingerprint Tanimoto, MCS Tanimoto, and MCS overlap coefficients. Candidate molecules with at least one similarity coefficient above an empirically determined threshold are selected.

4.5.2 Parameter optimization and performance

For each similarity measure, individual thresholds were determined during a 10-fold cross-validation among 623 known metabolites and 1 995 compounds of Recon. During the first fold means of the coefficient distributions of neighboring versus random pairs of known metabolites weighted by their interquartile ranges were estimated and used as first threshold: $WM_{fingerprint\,Tanimoto} = 0.74$, $WM_{\mathrm{MCS}\,Tanimoto} = 0.66$



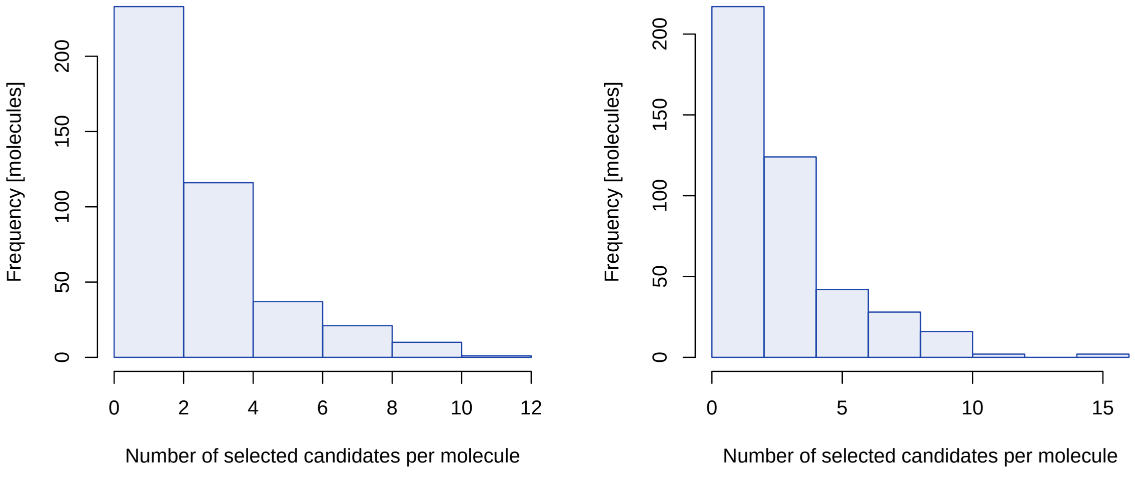
(a) Fingerprint Tanimoto (b) MCS Tanimoto (c) MCS Overlap

Figure 4.7: Distributions of coefficients during cross-validation
The maximum of fingerprint Tanimoto, MCS Tanimoto, and MCS overlap coefficients per metabolite of the test set and preselected database metabolites (Recon) show large differences for the neighboring metabolites respective random pairs. The blue lines indicate determined cutoffs according to version 2 (sensitivity weighted three-fold and specificity weighted once).

and $WM_{\text{MCS}\,overlap} = 0.91$ (Figure 4.7). Thresholds were systematically optimized over the following nine folds to maximize sensitivity and specificity (Table 4.7 and Table B.9). While specificity was weighted once, sensitivity was weighted twice (version 1) or three-fold (version 2). Parameter version 2 was maintained for following analyzes, because of an increased sensitivity and prediction rate, and still a reasonable number of selected candidates (Figure 4.8). The optimized parameter setting of 0.58, 0.66, and 0.9 for fingerprint Tanimoto, MCS Tanimoto, and MCS overlap, respectively, (version 2) leads to a sensitivity of 89% and a specificity of 55%.

Var.	Fingerp. Tanimoto	MCS Tanimoto	MCS Overlap	Sensitivity	Specificity	Rate	Freq.	Correct
1	0.68	0.66	0.9	0.86	0.62	67%	2.8	93%
2	0.58	0.66	0.9	0.89	0.55	69%	3.1	94%

Table 4.7: Empirically determined cutoffs for automated candidate selection
To determine cutoffs of variant 1 during the cross-validation sensitivity was weighted twice and of variant 2 sensitivity was weighted three-fold, while specificity was weighted once. Rate is the proportion of metabolites with at least one selected candidate molecule, followed by the average number of candidates per metabolite and the proportion of cases containing the correct candidate.



(a) Parameter variant 1

(b) Parameter variant 2

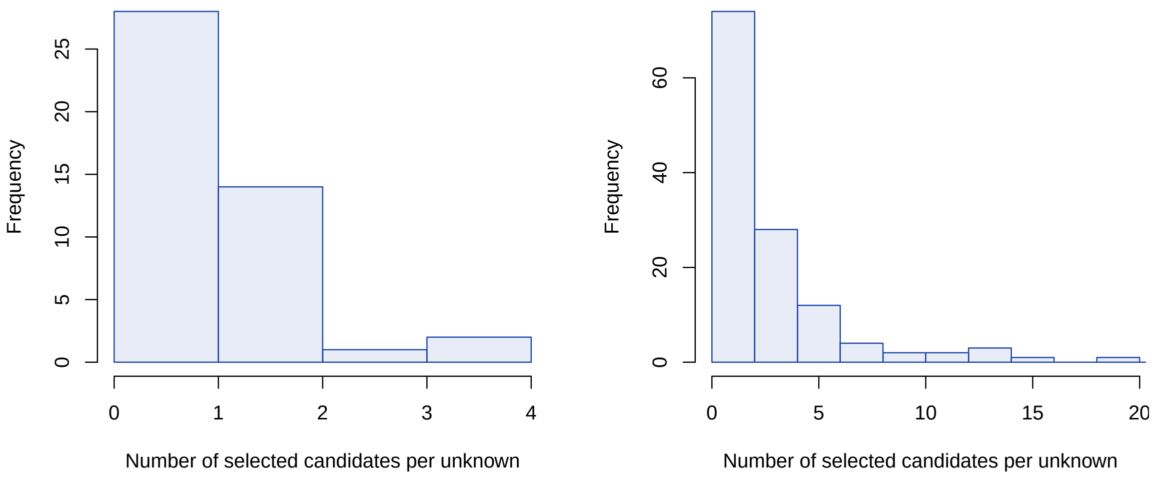
Figure 4.8: Number of selected candidates per molecule during validation
Both parameter variants lead to the selection of a maximum of four candidates for most molecules. These numbers strongly depend on the size and composition of the selection pool that contained 1 995 metabolites of Metabolon$^{\text{TM}}$ or Recon.

During the cross-validation among known metabolites 1 348 candidate molecules were automatically selected for 431 metabolites, corresponding to an average of

3.1 candidates for 69% of the compounds of the test sets. In 404 of 431 (93.7%) cases, the correct structure was in the set of automatically selected candidates.

4.5.3 Automatically selected candidates

During the first step of the automated procedure, we searched with measured masses of 282 (out of 315) unknown high-resolution metabolites that were connected to known metabolites with available structural information (fingerprint or molecular structure) in 1 995 compounds of Recon and 112 770 compounds of HMDB, and received 255 and 7 243 pre-selected candidate molecules for 132 and 274 unknown metabolites, respectively. After estimation of pairwise structural similarity (fingerprint Tanimoto, MCS Tanimoto, and MCS overlap scores) and filtering of the second and third step, 67 and 609 candidate molecules were automatically selected for 45 and 126 unknown metabolites based on Recon (Table 4.8, including evidences Table B.10) and HMDB (Table B.11 and Supplemental File B.8), respectively. In median, one candidate was selected per unknown metabolite within Recon and two candidates per unknown metabolite within HMDB (Figure 4.9). The maximum number of candidates per unknown metabolite is four for Recon and 106 for HMDB. Normalized



(a) Selection within Recon

(b) Selection within HMDB

Figure 4.9: Number of selected candidates per unknown metabolite
67 (a) and 609 (b) candidates for 45 and 126 unknown metabolites have been automatically selected within 1 995 compounds Recon and 112 770 additional compounds of HMDB, respectively. For the majority of unknowns, one to two candidates were selected based on Recon and one to six candidates were selected based on HMDB. In the second case three single unknowns with 30, 40 and 106 candidates were clipped for plotting.

Unknown	Candidate	PubChem CID	$\triangle$Mass	MFTC	MSTC	MSOC
X-02249	Benzo[a]pyrene-4,5-oxide	37786	0.0351	0.64*	0.48	0.85
	Benzo[a]pyrene-9,10-oxide	37456	0.0351	0.64*	0.47	0.77
X-11261	octenoyl carnitine	53481667	0.0068	0.61*	0.5	0.91*
X-11478	Perillic acid	1256	0.0073	0.61*	0.53	0.75
	(4-hydroxy-3-methoxyphenyl)acetaldehyde	151276	0.0291	0.6*	0.53	0.73
X-11787	L-4-hydroxyglutamic semialdehyde	25201126	0.0436	0.62*	0.64	0.9*
X-12093	L-alanyl-L-leucine	6992388	0.0071	0.94*	0.86*	1*
X-12100	5-hydroxy-L-tryptophan	439280	0.0071	0.67*	0.33	0.58
X-12170	trans-caffeate	689043	0.0185	0.58*	0.65	0.85
	3-Hydroxykynurenamine	440736	0.0291	0.84*	0.65	0.85
	5-Hydroxykynurenamine	164719	0.0291	0.79*	0.56	0.77
	adrenochrome o-semiquinone	10313383	0.0053	0.61*	0.5	0.69
X-12543	3-Methoxy-4-hydroxyphenylglycolaldehyde	440729	0.0069	0.77*	0.62	0.77
X-12688	L-alanyl-L-leucine	6992388	0.0073	0.69*	0.5	0.8
X-12860	glutaryl carnitine	53481622	0.0432	0.87*	0.85*	1*
X-13684	3-mercaptolactate-cysteine disulfide	193536	0.0068	0.73*	0.5	1*
X-13866	N-Ribosylnicotinamide	439924	0.018	0.66*	0.3	0.58
X-15497	fructoseglycine	3081391	0.0081	0.61*	0.5	0.69
	salsoline-1-carboxylate	320322	0.0071	0.61*	0.58	0.85
X-15503	dopachrome o-semiquinone	53481550	0.0309	0.62*	0.65	0.79
X-16071	6-amino-2-oxohexanoic acid	439954	0.0284	0.59*	0.56	0.9*
	L-allysine	207	0.0284	0.63*	0.47	0.8
	2-oxoglutaramate	48	0.008	0.6*	0.53	0.9*
X-16563	2-acetamido-5-oxopentanoate	192878	0.0295	0.83*	0.71*	1*
X-17323	16-hydroxypalmitate	7058075	0.0422	0.64*	0.32	0.58
X-18888	2-keto-3-deoxy-D-glycero-D-galactononic acid	22833524	0.0446	0.79*	0.57	0.8
X-21831	12-oxo-20-carboxy-leukotriene B4	53481457	0.021	0.7*	0.32	0.82
X-22158	hexanoyl carnitine	6426853	0.0068	1*	1*	1*
X-22379	5alpha-Dihydrotestosterone glucuronide	44263365	0.0063	0.96*	0.78*	0.88
	androsterone 3-glucosiduronic acid	114833	0.0063	1*	1*	1*
X-22836	6-amino-2-oxohexanoic acid	439954	0.0073	0.59*	0.5	0.7
	L-allysine	207	0.0073	0.64*	0.62	0.8
X-23196	5-L-Glutamyl-L-alanine	440103	0.0177	0.94*	0.6	0.8
	gama-L-glutamyl-L-alanine	11183554	0.0177	0.94*	0.6	0.8
	L-leucyl-L-serine	40489070	0.0187	0.73*	0.6	0.8
	N-acetylserotonin	903	0.0025	0.79*	0.48	0.69
X-23423	2-Amino-3-oxoadipate	440714	0.0289	0.65*	0.56	0.75
X-23581	1-piperideine-6-carboxylate	45266761	0.0073	0.57	0.67*	0.89
X-23583	D-proline	8988	0.0074	0.61*	0.6	0.75
	acetamidopropanal	5460495	0.0074	0.62*	0.45	0.62
X-23593	N-(omega)-Hydroxyarginine	440849	0.0043	0.61*	0.43	0.69
X-23647	Glycylproline	79101	0.0294	0.79*	0.75*	1*
	L-Prolinylglycine	6426709	0.0294	0.74*	0.75*	1*
X-23747	N1,N8-diacetylspermidine	389613	0.0074	0.98*	0.71*	1*
X-23776	6-amino-2-oxohexanoic acid	439954	0.0435	0.61*	0.67*	0.8
	L-allysine	207	0.0435	0.66*	0.67*	0.8

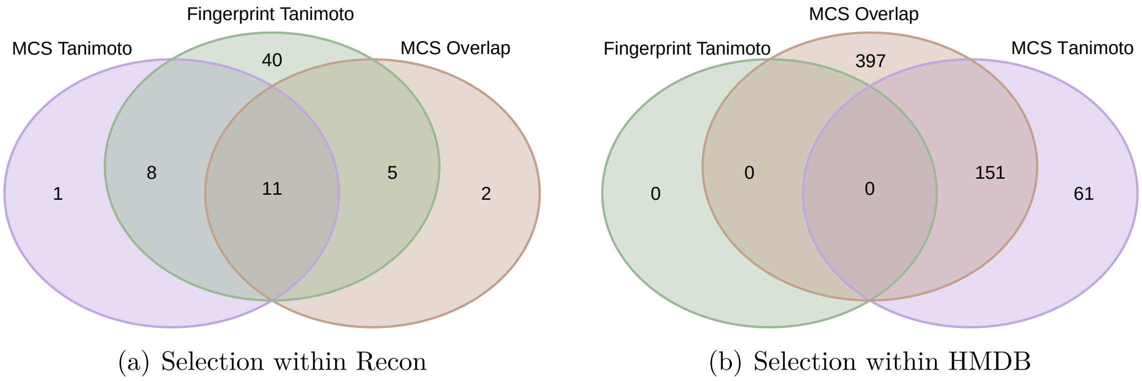
Table continued (p. 2/2)

Unknown	Candidate	PubChem CID	△Mass	MFTC	MSTC	MSOC
X-24330	Porphobilinogen	1021	0.0187	0.42	0.42	1*
X-24339	Xanthosine 5'-phosphate	73323	0.0336	0.94*	0.73*	0.92*
X-24361	16-Glucuronide-estriol	122281	0.0291	0.63*	0.65	0.79
	testosterone 3-glucosiduronic acid	108192	0.0073	0.84*	0.74*	0.85
X-24410	D-proline	8988	0.0072	0.62*	0.7*	0.88
	acetamidopropanal	5460495	0.0072	0.67*	0.67*	0.86
X-24411	D-proline	8988	0.0072	0.62*	0.64	0.88
	acetamidopropanal	5460495	0.0072	0.63*	0.5	0.75
X-24417	16-Glucuronide-estriol	122281	0.0287	0.63*	0.65	0.79
	testosterone 3-glucosiduronic acid	108192	0.0077	0.84*	0.74*	0.85
X-24452	3-Hydroxy-N6,N6,N6-trimethyl-L-lysine	439460	0.0071	0.81*	0.93*	1*
X-24455	N-acetyl-5-methoxykynuramine	390658	0.0294	0.71*	0.52	0.8
	N-formyl-L-kynurenine	910	0.007	1*	1*	1*
X-24498	cis-beta-D-Glucosyl-2-hydroxycinnamate	5316113	0.0078	0.64*	0.27	0.53
X-24514	putreanine	53477800	0.044	0.59*	0.35	0.55
X-24736	Glycylleucine	92843	0.0184	0.71*	0.5	0.73
	Leucylglycine	97364	0.0184	0.68*	0.41	0.64
X-24766	L-3-Cyanoalanine	439742	0.0437	0.76*	0.7*	0.88
X-24798	4-hydroxy-2-nonenal	5283344	0.0074	0.23	0.15	0.91*
X-24860	Biotin	171548	0.0466	0.58*	0.41	0.75
X-24983	D-argininium(1+)	71070	0.032	0.6*	0.57	0.8

Table 4.8: Selected candidate molecules based on Recon
67 candidate molecules have been automatically selected for 45 unknown metabolites based on fingerprint or structural similarity to neighboring known metabolites within the Recon database. The last three columns contain the maximum values of the fingerprint Tanimoto coefficient (MFTC), structure Tanimoto coefficient (MSTC), or structure overlap coefficient (MSOC) across pairs of the candidate molecule and known neighbors of the unknown metabolite. An asterisk indicates whether values are beyond the threshold. Detailed evidence information with names of neighboring known metabolites for each unknown metabolite are provided in Table B.10.

against the database sizes 3.4% of all Recon compounds and 0.5% of all HMDB compounds were selected as candidates.

Within compounds of Recon, 40 of 67 candidates passed the threshold of the fingerprint Tanimoto coefficient solely, and further 24 candidates additionally passed the thresholds of MCS Tanimoto or MCS overlap (Figure 4.10(a) and Table 4.8). 11 of those candidates passed the thresholds of all coefficients. In case of HMDB 548 of 609 candidates were selected by the MCS overlap criterion (Figure 4.10(b)). 151 of those and 61 further candidates were in addition or solely selected based on the MCS Tanimoto criterion. In this demonstration no HMDB-candidate was selected based on the fingerprint Tanimoto coefficient, since fingerprints were not available for these molecules. However, molecular structures can be used to calculate fingerprints. In total, 676 candidate molecules were automatically selected for 139 unknown metabolites.



(a) Selection within Recon

(b) Selection within HMDB

Figure 4.10: Number of selected candidates per selection criterion
Candidates were selected, if at least one of the selection criteria based on the fingerprint Tanimoto, MCS Tanimoto, or MCS Overlap coefficients were beyond the specific cutoffs. (a) 40 of 67 candidates within the Recon Database were selected solely and further 24 candidates in combination based on the fingerprint Tanimoto coefficient. 11 of those candidates passed the selection cutoff of all three criteria. (b) Within HMDB most candidates were selected based on the MCS overlap and MCS Tanimoto coefficients, since PubChem fingerprints were not available for most HMDB compounds.

4.5.4 Computer aided manual candidate selection

In cases, where unknown metabolites were not measured on a high-resolution platform, or where the candidate pool did not contain a suitable structure, automated annotations (of pathways, reactions, and the biochemical context) may assist manual selection of candidate molecules. Here, we were first focused on automated annotations, especially containing predictions of dehydrogenation (/reduction) reactions (passing Δmass and ΔRI criteria) for unknown metabolites measured by the established Metabolon^{TM} platform in samples of SUMMIT to select most promising candidates. As an example, super pathway ‘lipid’ and sub pathway ‘fatty acid, dicarboxylate’ were predicted for the unknown metabolite X-13891 based on the network model. In addition, our automated procedure assigns a dehydrogenation/ reduction reaction (passing the Δmass and ΔRI criteria) to the unknown metabolite and the neighbor dodecanedioic acid. Applying probable changes on the structure of dodecanedioic acid incorporating predicted information leads to the release of two hydrogen atoms in the carbon-strand and formation of a double bond: dodecenedioic acid (Figure 4.11). While positioning the double bound, 5 specific candidate molecules were selected for X-13891 (Table 4.9). Further exemplary candidate selection (for unknown metabolites X-11905 and X-11583) incorporating predicted annotations, reactions, and neighborships is prepared in Figure B.5. In total, we manually selected 39 candidate structures for 5 unknown metabolites (classified as lipid (fatty acid: dicarboxylate, medium chain, long chain) with a predicted dehydrogenation reaction), based on these principles (Table 4.9). The double bond cannot be determined by our approach alone as we do not use any information from fragmentation spectra. Thus, after exclusion of candidate structures that are measured as known metabolites on our metabolomics platform, 5 − 12 molecules remained as candidates for the 5 unknown metabolites. For 4 out of 5 candidates, at least one molecule was commercially available. We tested these candidate molecules by LC-MS for experimental verification (Chapter 4.6).

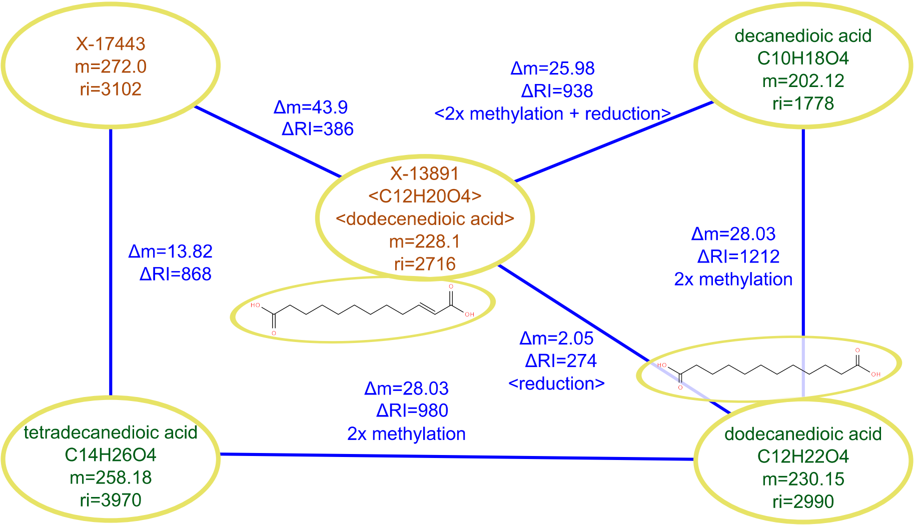


Figure 4.11: Neighboring metabolites of X-13891
Predicted super and sub pathways and predicted reactions with neighboring metabolites lead to the candidate molecule dodecenedioic acid. This candidate (with a certain double bound position) was experimentally verified later on. Multiple reactions, e.g. methylation + reduction, were added manually. Information in '<>' was predicted.

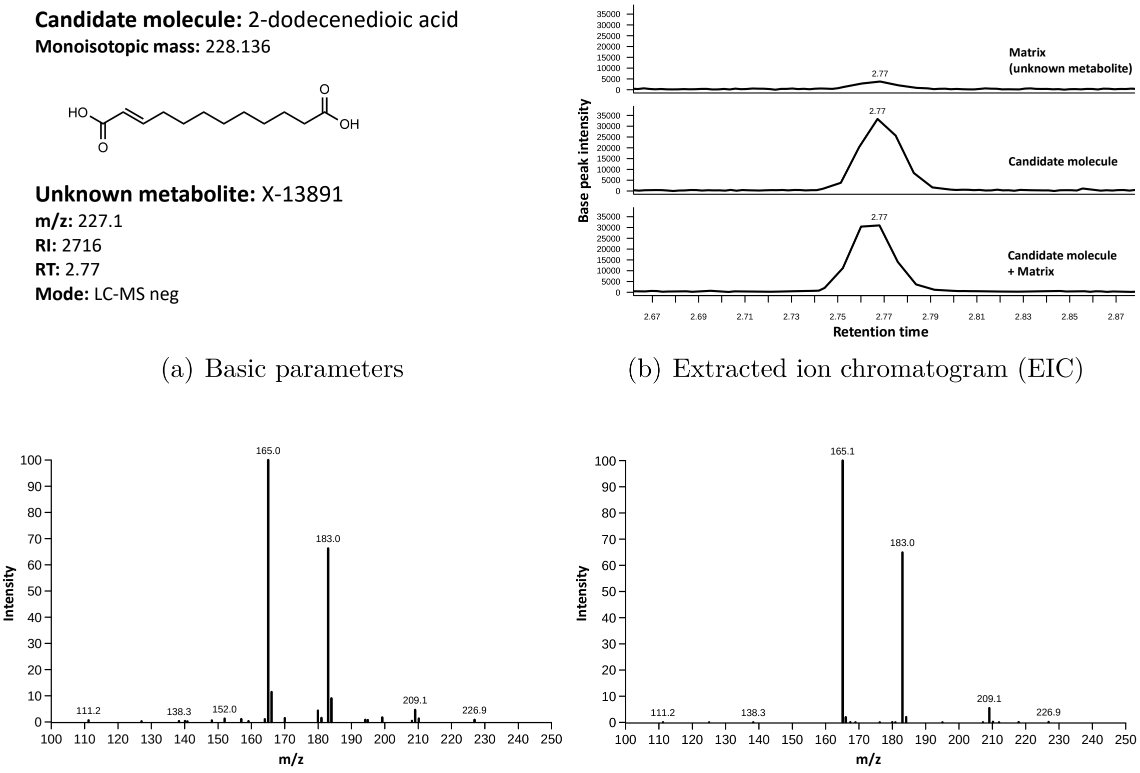
Metabolite	Super pathway	Sub pathway	Reaction	Reactant	Candidate molecules
X-13891	Lipid	Fatty acid, dicarboxylate	dehydro-genation	dodecanedioic acid	**2-dodecenedioic acid***, 3-dodecenedioic acid, 4-dodecenedioic acid, 5-dodecenedioic acid, 6-dodecenedioic acid
X-13069	Lipid	Long chain fatty acid	dehydro-genation	5,8-tetradeca-dienoate	2-tetradecenoic acid, 3-tetradecenoic acid, 4-tetradecenoic acid, 5-tetradecenoic acid, 6-tetradecenoic acid, 7-tetradecenoic acid, 8-tetradecenoic acid, **9-tetradecenoic acid***, 10-tetradecenoic acid, 11-tetradecenoic acid, 12-tetradecenoic acid, 13-tetradecenoic acid
X-11905	Lipid	Fatty acid, dicarboxylate	dehydro-genation	hexadecane-dioate	2-hexadecenedioic acid, 3-hexadecenedioic acid, 4-hexadecenedioic acid, 5-hexadecenedioic acid, 6-hexadecenedioic acid, 7-hexadecenedioic acid, 8-hexadecenedioic acid
X-11859	Lipid	Medium chain fatty acid	dehydro-genation	pelargonate	**2-nonenoic acid**, **3-nonenoic acid**, 4-nonenoic acid, 5-nonenoic acid, 6-nonenoic acid, 7-nonenoic acid, **8-nonenoic acid**
X-11538	Lipid	Fatty acid, dicarboxylate	dehydro-genation	octadecane-dioate	2-octadecenedioic acid, 3-octadecenedioic acid, 4-octadecenedioic acid, 5-octadecenedioic acid, 6-octadecenedioic acid, 7-octadecenedioic acid, 8-octadecenedioic acid, **9-octadecenedioic acid**
Procedure:	*automatic annotation*				*manual selection and verification*

Table 4.9: Automatically annotated and manually preselected candidate molecules Candidate molecules were selected manually for 5 unknown metabolites based on automatically predicted fatty acid derivatives and reactants in a dehydrogenation reaction. The structure of 5 − 12 candidate molecules per unknown metabolite basically varies in the position of the predicted double bond. Candidate molecules printed in bold were commercially available and those molecules marked by an asterisk were verified experimentally. I previously published a similar version of this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

4.6 Experimental verification of selected candidate molecules

Selected candidates usually pass the identification level 3, if the measured unknown metabolite and the candidate molecule (e.g. found in public databases) share common spectral, chemical, or physical properties. Before labels of unknown metabolites are replaced by the respective candidate names and are used for interpretations of scientific results, the candidate should be verified experimentally. Measurement of the pure substance may substantiate the prediction and leads to identification level 1 (verified). For a proof of concept, we sought experimental verification for candidates that were selected according to the most frequently predicted reaction type, the dehydrogenation reaction (Table 4.9). We bought 6 pure substances, predicted for 4 unknown metabolites, which were available at chemical distributors (Table 2.1). Applying LC-MS measurements (negative mode) with these pure substances, we were able to verify the predicted identity of two unknown metabolites.

2-dodecenedioic acid and X-13891 (m/z: 227.1) share a retention time (RT) peak at 2.77 min in their extracted ion chromatograms (EICs) and show the same fragments with equivalent relative intensities in their MS^2 fragmentation spectra, leading to a verification of this candidate molecule (Figure 4.12).



(a) Basic parameters

(b) Extracted ion chromatogram (EIC)

(c) MS^2 fragmentation spectrum: candidate

(d) MS^2 fragmentation spectrum: X-13891

Figure 4.12: Analytical comparison of 2-dodecenedioic acid and X-13891
The EIC of each aliquot shows the same retention time: candidate molecule, reference matrix (containing the unknown metabolite), and candidate molecule + reference matrix. Both MS^2 fragmentation spectra of the candidate molecule and the unknown metabolite show the same fragments with equal relative intensities leading to a verification of the candidate molecule. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Based on equal principles, we could verify one further candidate molecule: 9-tetradecenoic acid for X-13069 (Figure B.6). 9-octadecenedioic acid and X-11538 show a slight shift in their RT peaks so that this candidate molecule could be falsified, although both molecules show very similar fragments in their MS^2 fragmentation spectra (Figure B.7). 2-nonenoic acid, 3-noneoic acid, and 8-noneoic acid and the respective unknown metabolite X-11859 have different RT peaks in the EIC so that these candidates could also be falsified (Figure B.8). A detailed overview of the evaluation results (RT peaks and MS^2 signals) of all selected candidate molecules is provided in Table B.12.

4.7 Concepts to organize predicted or verified metabolite annotations

Our automated workflow for characterization of unknown metabolites expects a set of measured known and unknown metabolites (preferably equally sized) that is common for established non-targeted MS platforms such as Metabolon^{TM}. The network model can integrate several sets of independently measured metabolites and organizes characterized information, but it focuses on data analysis. Therefore, the transferability of automatically predicted or manually curated annotations of unknown metabolites measured in several cohorts requires an external organizational structure for storage that is independent form the characterization approach. Here, we propose a conceptual data structure and prototyped the database for storing annotation information of unknown metabolites including links to public databases for known or identified compounds (Figure 4.13).

In total, the database prototype contains 388 and 750 known, and 249 and 632 unknown metabolites measured on the established (SUMMIT cohort) or high-resolution (GCKD cohort) platform of Metabolon^{TM}, respectively. 59 unknown metabolites contain information of the measurement on both platforms. Predicted super and sub pathways are stored for 435 unknown metabolites and reactions connecting knowns to unknowns are stored for 384 unknown metabolites. 75 manually and 676 automatically selected candidate molecules are stored for 156 unknown metabolites. Finally, 2 candidates for formerly unknown metabolites are stored with identification level 1 including links to compounds of public databases, since selected candidates were verified experimentally (2-dodecenedioic acid for X-13891, and 9-tetradecenoic acid for X-13069). 831 of 984 measured known metabolites could be mapped to compounds of Recon, PubChem, or HMDB. Our suggested database aggregates annotations of overlapping sets of unknown metabolites that have been collected independently, and enables extraction of existing annotations for a given set of unknown metabolites occurring in results of novel experiments.

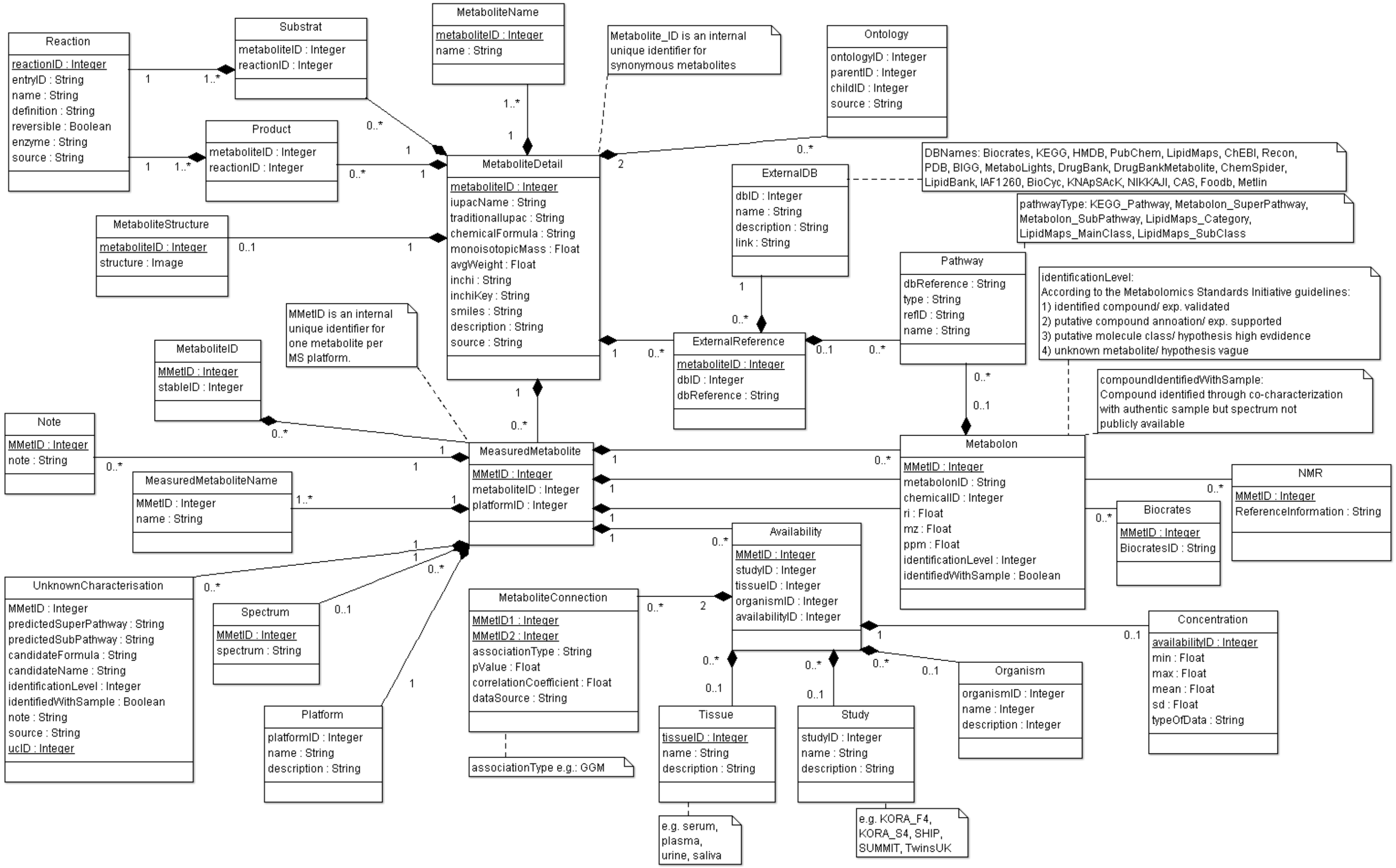


Figure 4.13: UML diagram of Metabolite Collection database
The Metabolite Collection database consists of two core parts: tables around 'MeasuredMetabolite' contain information regarding measurement, predictions, assumptions, verification, or identification of measured metabolites, whereas tables around 'MetaboliteDetail' contain information about molecules and molecule relations retrieved from public databases.

4.8 Discussion

Experimental identification of single unknown metabolites in wet-laboratories has gradually been superseded by complementary algorithm-based *in silico* approaches for performing high-throughput characterization of unknown metabolites in non-targeted MS-based metabolomics [127]. Here, we present a novel holistic method that integrates biochemical and genetic relations of unknown and known metabolites rather than their chemical or physical properties derived from spectra to annotate unknown metabolites and select candidate molecules. Our automated approach consists of independent modules that construct a systems biology model and access the network model for metabolite characterization: First, we identify the biochemical and genetic neighborhood of unknown metabolites that is imprinted in the correlation and association among measured metabolite levels and between these levels and the genotype of subjects in large cohorts. To this end, we construct the backbone of our network model, which already incorporates unknown and known metabolites, by pairwise partial correlations (GGM that are known to reconstruct biochemical pathways based on measured metabolite levels [147, 167]) and connect metabolites to genes based on genetic associations derived from a (published or performed) mGWAS. During the next step, our data integration module expands the data-driven network model by known biochemical reactions and metabolite-gene relations of the public database Recon [110]. While in principle other metabolic databases such as KEGG [38] or HumanCyc [115] can also be used as resources for previous biochemical information, interfaces or parsers need to be adapted for these databases.

The final network model is accessed by several automated metabolite characterization modules: We predict pathways by transferring annotations from known to unknown metabolites and assign reactions connecting known to unknown metabolites based on their local environment in the network model. Moreover, we automatically select isobaric candidate molecules for unknown metabolites (measured on high-resolution LC-MS devices) according to structural similarities of their neighbors with known structure (in the network model) to molecules of public databases, such as Recon [110] and HMDB [71].

To assess the quality of these predictions, we evaluated our approach based on the characterizations that our method produced for the set of metabolite pairs with known chemical structures. We demonstrated the flexibility of our approach with metabolite levels that were measured by a high-resolution MS platform in samples of a second independent cohort, replicated annotations within unknown metabolites (that could be mapped between both platforms), and showed that an integrated network of both data sets leads to annotations of additional unknown metabolites that could not be characterized separately. For verification of selected candidates, we tested several commercially available pure substances by LC-MS measurements in a wet-laboratory to search for experimental evidence. Finally, we proposed a database structure to organize generated metabolite annotations, to compare annotations of unknown metabolites between several data sets, and to characterize unknown metabolites of a new data set based on previous (predicted or verified) annotations.

4.8.1 Large scale metabolite characterization without fragment spectra

Our systems biology approach complements existing methods in various ways: By focusing on metabolite levels measured in samples of a cohort study, as well as genetic associations and functional relations of metabolites, we derive orthogonal information in our network model. This information is typically omitted in existing approaches that primarily rely on signals from fragmentation spectra (e.g. FingerID [90], CSI:FingerID [91, 92], SIMPLE [93], MetAssign [86], CFM [87], MetFrag [88], compMS^{2}Miner [89], and MetMapR [94]). Current approaches apply *in silico* fragmentation [87–89], calculation of RT [88, 89], or prediction of molecular fingerprints [90–93] to find candidate molecules with similar predicted structural properties based on compounds of public databases. Further concepts utilize networks that depict the similarity of unknown and known metabolites or signals in terms of detected fragments [94, 206], measured mass-to-charge ratios [94, 148], or elemental composition [207, 208].

As metabolomics measurements of large (e.g. epidemiological) cohort studies are regularly performed on standardized MS platforms by companies (such as Metabolon^{TM}) in a fee-for-service manner, there is usually no in depth spectral information provided that is needed for the described existing methods. In these cases, in which metabolite levels of a large number of samples are determined, our method provides an alternative approach for metabolite identification as it does not require spectral details beyond the reported quantities (cohort study), total mass-to-charge ratios, and RIs: On the one hand, our approach works independently of spectral information. On the other hand, our approach depends on the availability of measurements for a large number of samples, while methods that focus on spectral details usually only need the spectrum of the unknown metabolite from a single sample. Estimation of pairwise partial correlations (for GGMs construction) requires an at least balanced number of samples and measured parameters to receive robust relationships. Moreover, metabolite values need to be complete for all samples. For the most part, we solve this issue through removal of metabolites with more than 20% and samples with more than 10% missing values and imputation of remaining missing values. Nonetheless, unknown metabolites that show a very high number of missing values across samples cannot be annotated by our approach, since imputation is not applicable in these cases. Also, if a metabolite is not connected to any known metabolite or gene in our final network model, we cannot provide any further characterization for this unknown metabolite (besides assumed functional relationships to further unknown metabolites). In our case that only applied to a very small proportion of unknown metabolites. In classic epidemiology experiments, researchers are frequently interested in the identity (or at least in annotations) of a selected set of unknown metabolites that associate to certain biomarkers. These unknown metabolites can principally be integrated in our network by their associations to known components (metabolites or genes).

As its main strengths, our approach works independently of fragment spectra and identifies the biochemical environment of unknown metabolites, even if their molecular structure cannot be resolved in any case.

4.8.2 Complementary annotation through orthogonal information

By combining information that is imprinted in the correlation structure of metabolite levels and genetic associations with information extracted from spectral features such as mass-to-charge ratios, we were able to use mass differences between pairs of metabolites to characterize unknown metabolites even in MS data with low mass resolution. While mass resolution (even of our high-resolution platform) is insufficient to determine the atomic composition of an unknown metabolite, or to identify specific mass differences that are typical for certain reactions directly [148], our network allows pre-filtering these pairs by focusing on neighboring metabolites that can be assumed to be biochemically, i.e. functionally, linked.

As an integrated data-driven and knowledge-based network method, MetMapR [94] followed an idea that is similar to our approach. However, there are several differences: MetMapR starts with biochemical reactions between a user defined set of known metabolites that need to be available in or mapped to database structures and subsequently attaches unknown metabolites in a second or third step by spectral similarities or parametric Pearson or non-parametric Spearman correlations to (levels of) known metabolites. In part, the problem of missing database compounds is solved by adding hypothetical edges between known metabolites with similar PubChem fingerprints. In contrast, our approach profits from a direct inclusion of unknown metabolites by pairwise partial correlations that build a GGM, which is known to reconstruct biochemical pathways. Thus, most metabolites of the data set are directly included into the data-driven backbone of the network model and unknown metabolites are into their biochemically appropriate environment. While MetMapR applies different strategies to include known or unknown metabolites (the primary network contains known metabolites only; edges to unknown metabolites are added later on), the initial data-driven part of my network model includes known and unknown metabolites that were commonly integrated by the same principles, insuring a homogeneous and unbiased topology of known and unknown compounds. My approach is primarily focused on measured data, while database information is only used to support already existing structures in the network model. Metabolites that cannot be connected to parts of the previous network model will not be included to avoid unnecessarily large network substructures that cannot assist characterization of unknown metabolites. Although both methods, MetMapR and my approach, connect metabolites and genes according to common biochemical contexts, procedures, applied rules, and utilized information are different. The main focus of MetMapR is not (only) on identification of unknown metabolites but also on combining biochemical previous knowledge, structural similarity, and spectral similarity or reconstruct pathways, e.g. for diseases. In contrast to MetMapR, the main target of our approach consists of characterization of unknown metabolites that is performed by several automatized modules, while identification of health-, diseases, or environment related pathways are subordinate goals.

In general, our approach presupposes a similar behavior of unknown and known metabolites, which ignores potential biases in their distribution (e.g. if complete classes or pathways of metabolites are unknown). Moreover, in data sets where the number of unknown metabolites is much larger than the number of known metabolites, some unknown metabolites will only be connected to further unknown metabolites, which results in clusters of unknown metabolites. As a consequence, a transfer of pathway, reaction, and structural information to unknown molecules that are not connected to known compounds is limited.

Finally, while methods that do not rely on networks, are focused on the identification of a specific unknown metabolite, network-based methods such as our approach, provide annotations for most unknown metabolites in a metabolomics data set in a single run. Even metabolites, whose molecular structure cannot be elucidated, are characterized with information such as their biochemical environments. If new data sets from the same or a similar metabolomics platform become available, this data can easily be included to an existing network model. We showed that combined network models of several cohorts improve characterization of metabolites also for the previous data set.

4.8.3 Strategies to select specific candidate molecules

Most *in silico* approaches for automated selection of specific candidate molecules, including our method, search through a pool of molecular structures retrieved from public databases. However, searching criteria and strategies vary: While primary approaches compare measured spectra to spectra of databases [83–85], methods of next generations consider inter-peak dependency structures [86], or apply *in silico* fragmentation to compare observed and predicted spectra [87, 88]. Further methods (such as CSI:FingerID [90–92] or SIMPLE [93]) search public databases with PubChem fingerprints that were predicted based on measured fragment spectra of unknown metabolites. In contrast, our method avoids possible causes of errors or biases that may occur during prediction of fingerprints by searching with fingerprints of known metabolites that are located in the local environment of unknown metabolites according to the topology of our network model. We have shown, that metabolites that are connected in our network model are structurally more similar to each other than random pairs. Similarity of fingerprints is regularly estimated by the Tanimoto coefficient, which is used by CSI:FingerID and SIMPLE, and is also among the similarity measures of our approach. In general, fingerprints are appropriate to compare molecular structures, since they are fast to determine, common substructures can be identified, and the distance of fingerprints can be calculated easily and objectively.

While searching in a pool of structures, correct candidates can only be selected, if they are part of this pool. Therefore, a large candidate pool leads to an increasing number of selected candidates. With our method, we presuppose that considered reactions lead to minor structural changes of molecules, but the most similar structure is not necessarily the correct identity of an unknown metabolite in any case. However, we demonstrated the performance of our selection procedure based on fingerprint Tanimoto, MCS Tanimoto, and MCS overlap similarity scores in a set of known (measured) metabolites. The power of our metabolite selection module is in accessing the network model that is intended to connect functionally related and structurally similar metabolites as demonstrated in this study. Finally, our candidate selection module is only applicable for unknown metabolites measured on a high-resolution MS device to ensure an adequate pre-filtering by measured masses. Although (depending on the database) very specific molecules are selected, structurally similar (isotopic, isobaric, or stereo-isomeric) compounds should also be considered as molecules behind unknown metabolites. However, as one of our characterization modules, automatically selected candidates provide a solid idea about the possible biochemical identity of an unknown metabolite.

In general, automatically selected candidates should be evaluated manually or verified by experimental measurements, since automated selection approaches are restricted to compounds of a predefined set and to applied selection criteria that cannot capture the biochemically entirety that is in part modeled in our network. Even if the candidates have to be verified experimentally, the data driven selection of candidate molecules is a big gain, since a specific hint beyond the underlying molecule is provided for a large number of unknown metabolites and the manual work in the wet-laboratory is drastically reduced (compared to an experimental *de novo* identification of one single metabolite).

4.8.4 Future directions of metabolite characterization approaches

In future, advanced MS technologies that measure on the accurate mass resolution level will open the possibility to calculate the unambiguous chemical formula (elemental composition) based on the measured mass and possibly on a restricted set of atoms (which are frequently contained in biochemically active substances) [209]. Determined chemical formulas and predicted annotations will be used to search databases for a more precise filtering of candidate structures [210, 211].

Future inference of the sub- and super-pathways can potentially be based on less stringent independence assumptions. Although we receive a reasonable accuracy with our simplified approach of reaction prediction by solely relying on differences in the mass-to-charge ratio (and partially in the RI), this step can be improved by applying chemometric methods for *in silico* reaction prediction that take functional groups of the known metabolite into account [212]. In addition, the currently tested list of possible enzymatic reactions only covers the most frequent reactions and can be extended by multiple reactions (leading to an increase in the false positive rate and therefore requiring an advanced filtering).

4.8.5 Summary

Identification of metabolites measured on non-targeted MS-based platforms is among the major challenges of current metabolomics research [127]. Here we described a method that uses biochemical and genetic information as imprinted in the correlation structure of measured metabolite levels to characterize unknown metabolites. Integration of these metabolite-metabolite and metabolite-gene pairs with functional previous knowledge including enzymatic reactions and functional metabolite-gene relations of public databases in a network model embeds unknown metabolites into their metabolic contexts. Predicted pathways of unknown metabolites and reactions connecting known to unknown metabolites based on their local environment in the network model facilitates identification of unknown metabolites by both procedures, automated or manually. For unknown metabolites measured on high-resolution MS platforms, our approach automatically proposes specific candidate molecules of a database that are structurally similar to neighboring known metabolites within the network model. In future studies, metabolite identification will be largely improved by combining our approach with current spectra-based methods to incorporate complementary strategies.

CHAPTER 5

Summary and outlook

In the last decades, metabolomics has captured increasing importance in research of health and disease. The outcome of metabolomics analyses strongly depends on the quality of data produced by high-throughput technologies, such as mass spectrometry (MS) or nuclear magnetic resonance (NMR). Non-targeted approaches include measurements of metabolites whose chemical structures remain unknown and thus allow a hypothesis-free detection of unexpected relations. In MS, unknown metabolites are typically described by their analytical properties (e.g. m/z, retention time (RT), and features of fragmentation spectra), but lack any (bio-) chemical characterization. Even if measured metabolites carry some annotations, they are specified on specific precision levels that depend on the separation capacity of targeted as well as of non-targeted MS devices and on the analytical methodology used (e.g. multiple reaction monitoring (MRM) in tandem mass spectrometry (MS/MS)). In this regard, the total concentration of a mixture of several molecules may be measured and annotated as single metabolite. As an example, phosphatidylcholines (PCs) are often measured on the lipid species level, in which the total length and desaturation is determined, but their distribution to the *sn*-1 and *sn*-2 positions are not captured. Interpretation of scientific results that contain unknown or ambiguously annotated metabolites is limited. Frequently, these metabolites cannot be mapped to biochemical pathways or constituents of measured metabolite sums may have contrary biochemical functions. For biological interpretations and for further usage (e.g. clinical applications), these metabolites need to be characterized precisely.

5.1 Achieved goals in metabolite characterization

In this work, I introduced two automated metabolite characterization methods: In the first part, a comparison of two targeted MS platforms led to uncovering the main constituent on the fatty acid level (detection of fatty acid chain lengths and desaturation, but not their order (*sn*-1 or *sn*-2), position of double bounds, and stereo-chemistry) for PCs measured on the lipid species level. To this end, I first systematically divided chain lengths and double bonds of 76 PCs measured by Absolute*IDQ*$^{\text{TM}}$ (lipid species level) between their *sn*-1 and *sn*-2 positions and received 150 to 456 theoretically possible constituents of the fatty acid level. For a subset of 38 PCs sums (lipid species level) at least one constituent was measured among 109 PCs of the fatty acid level (Lipidyzer$^{\text{TM}}$) that were quantified in the same 223 human plasma samples of HuMet. For 10 PCs measured on the lipid species level, I identified one single main constituent with a relative contribution of at least 80% in the sum. Further 14 and 3 PC sums consisted of one or two measured constituents, respectively, with a contribution of $20-80\%$. Finally, I demonstrated the stability of the quantitative contribution of constituents through replication in 305 samples of an independent cohort (QMDiab): 13 of 38 quantitative PC compositions by comparing their estimated contribution directly (factors f) and based on measured and imputed concentrations (fatty acid level) 19 PC compositions. Determined stable factors can thus be used to impute PC concentrations (fatty acid level) of further studies based on PC concentrations measured on the lipid species level (e.g. Absolute*IDQ*$^{\text{TM}}$).

In the second part, I generated a systems biology model that integrated data-driven and database relations between characterized and uncharacterized metabolites in non-targeted MS). By transferring annotations from known to neighboring unknown metabolites, I identified the biochemical context of unknown metabolites. For this purpose, I estimated pairwise partial correlations between 637 metabolites (338 known, 249 unknown) measured by the established platform of Metabolon$^{\text{TM}}$ in 2 279 samples of SUMMIT, since resulting interactions are known to reconstruct biochemical pathways [118]. The resulting GGM (862 edges) was combined with 398 published partial correlations of metabolites (KORA) [147] and 299 metabolite-gene associations of a published mGWAS (KORA and TwinsUK) [167]. Thereby, common edges were collected, common metabolites were mapped, and new metabolites were added. Finally, I attached 306 metabolites and 57 genes connected via 1 152 reactions and 495 metabolite-gene relations from the public database Recon [110] to the data-driven network model. The resulting network model was the basis for several metabolite characterization modules: (*i*) 180 unknown metabolites were labeled with predicted pathways that were transferred from metabolites in the local biochemical environment according to the network model. (*ii*) Mass differences that were typical for 21 reaction types were used to assign 100 potential enzymatic reactions to pairs of neighboring (acc. to the network model) known and unknown metabolites. In case of mass difference 2, I applied an additional ΔRI criterion and thus predicted dehydrogenation reactions for 12 of the 100 pairs. (*iii*) For 139 out of 315 unknown metabolites that were measured on the new high-resolution platform of Metabolon$^{\text{TM}}$ in 1 209 samples of a second cohort (GCKD) and were connected to known metabolites in the data-driven network model, I automatically selected 676 candidate molecules by scanning compounds of Recon [110] and HMDB [71] for similar masses of unknown metabolites and similar chemical structures to their local environment in the network model. (*iv*) Finally, I prototyped a database structure for storing (predicted, verified, or known) annotations of unknown metabolites.

To demonstrate applications of my automated workflow, I manually selected 5 – 12 candidate molecules for 5 unknown metabolites based on described automatic annotations, of which I verified two candidates experimentally in a wet-laboratory. For replication of automatic metabolite characterizations, I mapped 59 of 632 unknown metabolites that were measured in the second cohort to metabolites of the main data set (SUMMIT). Thus, I replicated pathway predictions for 18 out of 26 unknown metabolites that carried predictions in both cohorts and I showed that predictions can be improved by integration of two data sets into a common network model (enabled pathway predictions for another 8 unknown metabolites). The described systems biology approach covers automated steps from data integration to annotation of pathways, prediction of reactions, and selection of specific candidates, while reducing manual work to a minimum.

5.2 Biological benefit of precise metabolite annotations

Non-targeted MS technologies are established label-free approaches to quantify metabolites in easily extractable tissues (e.g. blood plasma, serum, urine, or saliva) and to explore unexpected relations without specific hypotheses. Current epidemiology studies (e.g. association analyses, mGWAS, or MWAS) frequently exclude the subset of unknown metabolites from experiments [34], or report results including unknown metabolites without further interpretations [167]. Only a few studies search for hints concerning the identity of unknown metabolites [213]. Uncharacterized metabolites in research results mark dead ends with a limited scientific value. Annotations enable interpretations and illumination of the biochemical background of results that contain unknown metabolites. If unknown metabolites should, as an example, serve as biomarkers they need to be identified for exploration of functional relations to specific traits (e.g. diseases).

Besides unknown metabolites, some annotated metabolites carry ambiguous labels, since they consist of several chemical structures that have similar chemical or physical properties according to the separation capacity of the MS device or of the identification analysis (precision level). To enable interpretations of results containing these metabolites, they often use assumptions regarding the underlying chemical structure (e.g. even chain lengths are much more abundant than odd chain lengths [138]; especially C16:0, C18:0, C18:1, C18:2, and C20:4 are common fatty acids in human plasma [139]). Based on the platform comparison described in this work, I could partially confirm or falsify respective expectations in published studies: as an example, Suhre et al. 2011 supposed that PC aa C36:4, PC aa C38:4, PC aa C38:5 and PC ae C38:3 consist of one saturated or mono-unsaturated fatty acid with chain length C16 or C18 and of one poly-unsaturated fatty acid [140], which corresponds to the results of my analyses. As a second example, Floegel et al. 2013 assumed PC 14:0_18:1 as major constituent of PC aa C32:1 [141], but my data showed that this compound is only among its minor constituents. Empirically determined quantitative PC compositions (fatty acid level) of my study are still not on the level of chemical structures. However, they enable more substantiated interpretations by reducing the number of possible molecules.

Especially in clinical applications metabolomics testing requires precisely characterized metabolites, since missing or ambiguous annotations are not sufficient to find specific disease causes, to functionally explore potential treatment targets, to pass analytical and clinical validation, or to meet regulatory frameworks [214]. However, scientific interpretations of results can be supported by placing unknown metabolites in their biochemical environment and label them with functional elements, such as pathways. While existing spectra-based metabolite annotation approaches try to identify (parts of) metabolite structures, my approach supplies the local biochemical environment for a large set of unknown metabolites, even for metabolites whose identity cannot be resolved. Resulting biochemical context information about unknown metabolites is a significant improvement on pure analytical data (m/z, RT, or spectra).

On the downside, data-driven construction of systems biology models that depict biochemical systems including interconnected elements of various layers (e.g. metabolites, enzymes, genes, tissues, or external factors) require large data sets. In small intervention (or case-control) studies, data is often not sufficient to derive an appropriate biochemical model and special interest is in identification of particularly selected singular unknown metabolites. In these experimental settings, metabolite identification is preferably focused on fragment spectra (e.g. FingerID [90], CSI:FingerID [91, 92], SIMPLE [93], MetAssign [86], CFM [87], MetFrag [88], compMS^{2}Miner [89], and MetMapR [94]).

Whenever the data basis is sufficient to model biochemical relations incorporating unknown metabolites, metabolite characterization should utilize this valuable information. In my approach, I construct and access data-driven network models (reconstructing biochemical pathways) to identify the biochemical context of unknown metabolites, transfer pathway information from known to neighboring unknown metabolites, predict reactions between known and unknown compounds, and automatically select specific candidate molecules. Automatically selected information concerning unknown metabolites helps to understand the biochemical mechanisms behind computational results and enables scientific interpretations.

Precisely characterized metabolites are crucial for reliable interpretations of metabolomics study-results, especially if they are intended for later clinical applications. Depending on the available data and on the specific use case, it might be an advantage to combine several metabolite characterization approaches (based on fragment spectra and based on the biochemical environment).

5.3 Technical aspects on automated characterization approaches

Experimental metabolite identification approaches in wet-laboratories (by non-targeted MS) are successfully applied to elucidate the structure of single unknown metabolites. However, these approaches are comparatively expensive, time-consuming, and require large probe volumes (because several measurements are necessary). While current non-targeted MS platforms measure several hundred or thousand metabolites, upcoming technologies are becoming faster and faster, leading to an increased throughput of samples and measurement of a growing number of (unknown) metabolites (as a comparison, HMDB contains more than 100 000 metabolites that are presumed to be biochemically active in human metabolism [71]). In theory, all compounds of human metabolite databases (such as HMDB) could be sent through an MS measurement of a certain platform and stored in a spectral library to reduce the number of unknown metabolites. In practice, this concept is not feasible, since the number of database compounds is very large, only a small subset of these compounds is commercially available, detected unknown metabolites are not necessarily found in public databases, synthesis of specific compounds (as fee-for-service manner) is very expensive, once identified compounds cannot necessarily be transferred to further platforms, and the procedure would not be fundable for medium-sized MS companies. As a consequence, current non-targeted metabolomics requires high-throughput approaches that target several unknown metabolites simultaneously.

Automated *in silico* approaches usually transfer information from known to unknown entities or predict physical as well as chemical properties. While most approaches are driven by fragment spectra [83–94], some methods, such as my systems biology model, identify biochemical contexts of unknown metabolites [94, 150]. The most basic methods compare complete measured MS or MS/MS spectra of an unknown metabolite to reference spectra of databases (e.g. MassBank [215] or METLIN [216]), in which availability and transferability of spectral information depend on the measuring device [83–85]. This procedure can be improved by clustering m/z and intensity of all peaks of unknown metabolites and target molecules of spectral databases separately and ranking similar structures (MetAssign [86]). In order to become independent of spectral databases, several approaches apply *in silico* fragmentation of a predefined set of chemical structures (in part supported by fragmentation trees) and compare resulting predicted spectra to measured spectra of unknown metabolites to select candidate molecules (CFM [87], MetFrag [88], and compMS^{2}Miner [89]). Besides physical properties (e.g. fragmentation spectra), some methods filter candidate molecules by (partially predicted) chemical properties, such as RT (MetAssign [86], MetFrag [88], compMS^{2}Miner [89], and my approach [150]). Kernel-based machine learning approaches predict chemical fingerprints (e.g. PubChem fingerprint consists of a 881 bit tuple) based on fragmentation spectra, and query databases to find candidate molecules that show a similar fingerprint (FingerID [90], CSI:FingerID [91, 92], MetMapR [94], and SIMPLE [93]). As an example, similarity of chemical fingerprints can be estimated by the Tanimoto coefficient. In my approach, automated candidate selection also employs chemical fingerprints, in which prediction of fingerprints is not necessary, since fingerprints of known metabolites located in the local biochemical environment of unknown metabolites are used. All of these approaches, including my approach, have in common that candidates can only be selected for unknown metabolites, if the candidate is among accessed database compounds. Unlike approaches that are restricted to chemical or physical properties of unknown metabolites, systems biology methods characterize unknown metabolites by their biochemical context, even if no candidate is predicted (MetMapR [94] and my approach [150]). Spectra-based approaches are independent from metabolomics experiments and require only the measurement of one single sample. Admittedly, spectral information must be available, and methods do not provide annotations of pathways or biochemical relations to further metabolites. Some commercially available non-targeted MS platforms (such as Metabolon^{TM}) that are regularly used for standardized metabolite quantification in large cohorts do not provide spectral information (except the total m/z). If these data sets are used for large-scale hypothesis-free testing (e.g. MWAS or mGWAS), resulting associations can serve for metabolite identification by systems biology approaches that use correlation structures to place unknown metabolites into the correct biochemical context.

My approach is focused on transfer of orthogonal metabolite annotations based on network models that incorporate known and unknown metabolites connected by association or functional information. Data-driven elements, such as partial correlations between pairwise metabolites that reconstruct biochemical pathways incorporating unknown metabolites (GGM) built the backbone of my model and are integrated with metabolite-gene associations (mGWAS), whereas GGM generation requires at least balanced numbers of samples and variables (metabolites). Therefore, my approach is suitable for characterization of unknown metabolites that appear in results of large cohort studies. The resulting network model is revalued with reactions and functional metabolite-gene relations of a public database (e.g. Recon [110]) and provides a basis for flexible metabolite characterization modules: Pathway annotations are transferred from known to neighboring unknown metabolites and reactions are predetermined by co-localization in the network model and characteristic mass changes. For unknown metabolites with measured high-resolution masses, candidate molecules are automatically selected from compounds of a public database (e.g. Recon, HMDB) that are structurally similar to metabolites located in the biochemical context (acc. to the network model) and meet the measured mass of the unknown metabolite. The success of my approach strongly depends on the availability of large data sets (e.g. cohort studies), in which fractions (e.g. half) of measured metabolites are known or already identified.

Besides emerging metabolite characterization software, further development of MS devices leads to an improved metabolite separation capacity (better precision level) (e.g. SelexION$^{\text{TM}}$ technology) and to an increasing mass resolution (high-resolution MS): First, current lipidomics platforms (such as Lipidyzer$^{\text{TM}}$) quantify several thousand metabolites on the fatty acid (acyl/alkyl) level, but many data sets of large epidemiological cohort studies have already been measured by previous platforms meeting the lipid species level (e.g. Absolute*IDQ*$^{\text{TM}}$). While using existing data in scientific analyses, further knowledge about their probable composition (main constituents on a higher precision level), as determined in my study, is useful for satisfying interpretations of respective results. Data-driven predictions of constituents decrease the necessity of remeasuring large numbers of samples on new devices, which is not possible in any case, if samples are already depleted or reserved for further analyses. Although current platforms that measure on the fatty acid level apply a much better separation of metabolites, they are far from measuring the chemical structure (e.g. in case of fatty acids, position of double bonds, stereo-chemistry, etc. is not detected). Therefore, prediction of probable constituents will still be necessary (even if platforms with even better precision levels are developed).

The second aspect of evolving MS devices is about the measured mass resolution. Detection of the accurate mass enables determination of the atomic composition of molecules, in which precision decreases with the measured mass (e.g. $m_1 = 100 \pm 0.01$ and $m_2 = 1\,000 \pm 0.1$) and the number of combinatorically possible atomic compositions increases exponentially (with the number of atoms) for large metabolites. Developing MS technologies (high-resolution measurements or better metabolite separation capacity) require adaptation of metabolite characterization approaches, but *in silico* metabolite characterization cannot be replaced by recent MS devices yet.

Recent MS platforms usually cover more extensive sets of known and unknown metabolites (e.g. established Metabolon$^{\text{TM}}$ platform: ~ 750 metabolites, new high-resolution Metabolon$^{\text{TM}}$ platform: $\sim 1\,450$ metabolites). Large numbers of unknown metabolites require high-throughput approaches that annotate a complete set of unknown metabolites during a single run. Especially metabolite characterization procedures that focus the local biochemical environment of unknown metabolites (instead of fragment spectra), such as my approach, profit from large numbers of samples and of metabolites, since annotations can be transferred from (more) known to unknown metabolites via an increased number of neighborhood relations.

In summary, available information of the measurement (number of samples, parameters: m/z, RT or RI, spectra) influences selection of the optimal metabolite characterization approach. Unknown metabolites in data sets with few or only one sample should be approached by methods that focus fragment spectra, if spectra are available. In cases, where spectra are not available, but metabolite concentrations are measured in large cohort studies, procedures focusing the biochemical context of unknown metabolites, such as my method, are suitable. In this case, metabolite characterization is approached by the top down principle, in which information is retrieved from external overlying factors, such as associations or an orthogonal transfer from known entities.

5.4 Future directions in handling structurally uncharacterized metabolites

Novel targeted and non-targeted MS technologies are developed repeatedly according to the evolving state of the art. Characterization of unknown metabolites needs to follow this development. First, more efficient MS technologies realize fast scans during measurements of large cohort studies within short time spans, e.g. German NAKO ($n \approx 200\,000$) [24]. My metabolite characterization approach benefits from large sample sizes, since estimation of pairwise partial correlations (for GGM generation) requires an at least balanced number of variables and instances. In general, the statistical power rises by increasing sample sizes, meaning that associations with smaller effect sizes are detected. Second, (most of) recently developed MS technologies achieve a higher mass resolution, compared to established platforms, and often allow a better separation of molecules. As an example, the lipidomics platform Lipidyzer$^{\text{TM}}$ quantifies more than a thousand metabolites on the fatty acid level, leading to a direct analytical selection of metabolites that my approach predicted based on metabolites measured on the lipid species level. Although measurements on higher precision levels open new possibilities, the separation capacity is still far from detecting actual chemical structures and therefore justify determination of their constituents. Furthermore, emerging high-resolution MS technologies implement precise measurement of metabolite masses. In my approach, high-resolution masses serve as an efficient filtering criterion for pre-selection of candidate molecules that are structurally similar to compounds located in the biochemical context of unknown metabolites. Measurements of accurate masses even allow the determination of unique atomic compositions of metabolites.

Future trends in metabolite identification will further shift from manual to efficient *in silico* approaches that allow high throughputs of large data sets. Limiting factors, i.e. manual tasks, are systematically reduced by introducing automated modules. Several methods employing specific data (e.g. my approach transfers annotations from known to unknown metabolites by identification of the biochemical context; other methods, such as FingerID [90], CSI:FingerID [91, 92], SIMPLE [93], MetAssign [86], CFM [87], MetFrag [88], compMS^{2}Miner [89], or MetMapR [94] utilize fragmentation spectra) will be combined to attain the most precise annotation for each unknown metabolite. Metabolite identification in wet-laboratories will be focused on single unknown metabolites of special interest (e.g. promising diagnostic or therapeutic targets) that cannot be annotated by automated approaches (e.g. if they are located in a cluster of further unknowns only, or if no fragment can be labeled). However, experimental measurements in wet-laboratories remain important for verification of (automatically) selected candidate molecules to reach identification level 1.

Results of metabolomics experiments can only be understood, placed into biochemical contexts, interpreted, or further used in clinical applications, if relevant metabolites are annotated with their chemical structure. Since novel MS platforms, facing the actual state of technology, are developed, data sets containing new (yet unobserved) unknown metabolites are produced. For this reason, metabolite identification will stay a long-term companion in future metabolomics research [217]. In case of more and more emerging non-targeted MS platforms, the focus may change from *de novo* identification of unknown metabolites to comparative approaches that transfer annotations between known and unknown metabolites measured on different platforms. With the ability of integrating metabolites of various platforms and illuminating data-driven biochemical contexts of unknown metabolites, my approach is well prepared for this future challenge.

Bibliography

[1] Danielle Ryan and Kevin Robards. Metabolomics: The greatest omics of them all? *Analytical chemistry*, 78:7954–7958, 2006. doi:10.1021/ac0614341.

[2] Igor Stagljar. Editorial for “advances in OMICs-based disciplines”. *Biochemical and biophysical research communications*, 445:681–682, 2014. doi:10.1016/j.bbrc.2014.03.040.

[3] Andrea D Weston and Leroy Hood. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of proteome research*, 3:179–196, 2004. doi:10.1021/pr0499693.

[4] Ina Aretz and David Meierhofer. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *International journal of molecular sciences*, 17(5):632, 2016. doi:10.3390/ijms17050632.

[5] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359, 2018. doi:10.1126/science.aao0185.

[6] Noha A Yousri, Gabi Kastenmüller, Christian Gieger, So-Youn Shin, Idil Erte, Cristina Menni, Annette Peters, Christa Meisinger, Robert P Mohney, Thomas Illig, Jerzy Adamski, Nicole Soranzo, Tim D Spector, and Karsten Suhre. Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics : Official journal of the Metabolomic Society*, 10:1005–1017, 2014. doi:10.1007/s11306-014-0629-y.

[7] M E Lacruz, A Kluttig, D Tiller, D Medenwald, I Giegling, D Rujescu, C Prehn, J Adamski, K H Greiser, and G Kastenmüller. Instability of personal human

metabotype is linked to all-cause mortality. *Scientific reports*, 8:9810, 2018. doi:10.1038/s41598-018-27958-1.

[8] Simone Rochfort. Metabolomics reviewed: a new “omics” platform technology for systems biology and implications for natural products research. *Journal of natural products*, 68:1813–1820, 2005. doi:10.1021/np050255w.

[9] Kedir N Turi, Lindsey Romick-Rosendale, Kelli K Ryckman, and Tina V Hartert. A review of metabolomics approaches and their application in identifying causal pathways of childhood asthma. *The Journal of allergy and clinical immunology*, 141:1191–1201, 2018. doi:10.1016/j.jaci.2017.04.021.

[10] Drupad K Trivedi, Katherine A Hollywood, and Royston Goodacre. Metabolomics for the masses: The future of metabolomics in a personalized world. *New horizons in translational medicine*, 3:294–305, 2017. doi:10.1016/j.nhtm.2017.06.001.

[11] Christopher B Newgard. Metabolomics and Metabolic Diseases: Where Do We Stand? *Cell metabolism*, 25:43–56, 2017. doi:10.1016/j.cmet.2016.09.018.

[12] Ioanna Tzoulaki, Timothy M D Ebbels, Ana Valdes, Paul Elliott, and John P A Ioannidis. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *American journal of epidemiology*, 180:129–139, 2014. doi:10.1093/aje/kwu143.

[13] Marianne C Walsh, Lorraine Brennan, J Paul G Malthouse, Helen M Roche, and Michael J Gibney. Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans. *The American journal of clinical nutrition*, 84:531–539, 2006. doi:10.1093/ajcn/84.3.531.

[14] J A Carrillo, M Christensen, S I Ramos, C Alm, M L Dahl, J Benitez, and L Bertilsson. Evaluation of caffeine as an in vivo probe for CYP1A2 using measurements in plasma, saliva, and urine. *Therapeutic drug monitoring*, 22: 409–417, 2000.

[15] Mónica Calderón-Santiago, Feliciano Priego-Capote, Natacha Turck, Xavier Robin, Bernabé Jurado-Gámez, Jean C Sanchez, and María D Luque de Castro. Human sweat metabolomics for lung cancer screening. *Analytical and bioanalytical chemistry*, 407:5381–5392, 2015. doi:10.1007/s00216-015-8700-8.

[16] R Holle, M Happich, H Löwel, H E Wichmann, and MONICA/KORA Study Group. KORA–a research platform for population based health research. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany))*, 67 Suppl 1:S19–S25, 2005. doi:10.1055/s-2005-858235.

[17] Wolfgang Rathmann, Bernd Kowall, Teresa Tamayo, Guido Giani, Rolf Holle, Barbara Thorand, Margit Heier, Cornelia Huth, and Christa Meisinger. Hemoglobin A1c and glucose criteria identify different subjects as having type 2 diabetes in middle-aged and older populations: the KORA S4/F4 Study. *Annals of medicine*, 44:170–177, 2012. doi:10.3109/07853890.2010.531759.

[18] Kai-Uwe Eckardt, Barbara Bärthlein, Seema Baid-Agrawal, Andreas Beck, Martin Busch, Frank Eitner, Arif B Ekici, Jürgen Floege, Olaf Gefeller, Hermann Haller, Robert Hilge, Karl F Hilgers, Jan T Kielstein, Vera Krane, Anna Köttgen, et al. The German Chronic Kidney Disease (GCKD) study: design and methods. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, 27:1454–1460, 2012. doi:10.1093/ndt/gfr456.

[19] Katja Borodulin, Erkki Vartiainen, Markku Peltonen, Pekka Jousilahti, Anne Juolevi, Tiina Laatikainen, Satu Männistö, Veikko Salomaa, Jouko Sundvall, and Pekka Puska. Forty-year trends in cardiovascular risk factors in Finland. *European journal of public health*, 25:539–546, 2015. doi:10.1093/eurpub/cku174.

[20] Ewan R Pearson, Louise A Donnelly, Charlotte Kimber, Adrian Whitley, Alex S F Doney, Mark I McCarthy, Andrew T Hattersley, Andrew D Morris, and Colin N A Palmer. Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. *Diabetes*, 56:2178–2182, 2007. doi:10.2337/db07-0440.

[21] Damiano Baldassarre, Kristiina Nyyssönen, Rainer Rauramaa, Ulf de Faire, Anders Hamsten, Andries J Smit, Elmo Mannarino, Steve E Humphries, Philippe Giral, Enzo Grossi, Fabrizio Veglia, Rodolfo Paoletti, Elena Tremoli, and IMPROVE study group. Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study. *European heart journal*, 31:614–622, 2010. doi:10.1093/eurheartj/ehp496.

[22] Axel C Carlsson, Per E Wändell, Gunilla Journath, Ulf de Faire, and Mai-Lis Hellénius. Factors associated with uncontrolled hypertension and cardiovascular risk in hypertensive 60-year-old men and women–a population-based study. *Hypertension research : official journal of the Japanese Society of Hypertension*, 32:780–785, 2009. doi:10.1038/hr.2009.94.

[23] T S Ahluwalia, E Lindholm, and L C Groop. Common variants in CNDP1 and CNDP2, and risk of nephropathy in type 2 diabetes. *Diabetologia*, 54: 2295–2302, 2011. doi:10.1007/s00125-011-2178-5.

[24] German National Cohort (GNC) Consortium. The German National Cohort: aims, study design and organization. *European journal of epidemiology*, 29: 371–382, 2014. doi:10.1007/s10654-014-9890-7.

[25] Susanne Krug, Gabi Kastenmüller, Ferdinand Stückler, Manuela J Rist, Thomas Skurk, Manuela Sailer, Johannes Raffler, Werner Römisch-Margl, Jerzy Adamski, Cornelia Prehn, Thomas Frank, Karl-Heinz Engel, Thomas Hofmann, Burkhard Luy, Ralf Zimmermann, et al. The dynamic range of the human metabolome revealed by challenges. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 26:2607–2619, 2012. doi:10.1096/fj.11-198093.

[26] Christian Gieger, Ludwig Geistlinger, Elisabeth Altmaier, Martin Hrabé de Angelis, Florian Kronenberg, Thomas Meitinger, Hans-Werner Mewes, H-Erich Wichmann, Klaus M Weinberger, Jerzy Adamski, Thomas Illig, and Karsten Suhre. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics*, 4:e1000282, 2008. doi:10.1371/journal.pgen.1000282.

[27] Thomas Illig, Christian Gieger, Guangju Zhai, Werner Römisch-Margl, Rui Wang-Sattler, Cornelia Prehn, Elisabeth Altmaier, Gabi Kastenmüller, Bernet S Kato, Hans-Werner Mewes, Thomas Meitinger, Martin Hrabé de Angelis, Florian Kronenberg, Nicole Soranzo, H-Erich Wichmann, et al. A genome-wide perspective of genetic variation in human metabolism. *Nature genetics*, 42: 137–141, 2010. doi:10.1038/ng.507.

[28] Gabi Kastenmüller, Johannes Raffler, Christian Gieger, and Karsten Suhre. Genetics of human metabolism: an update. *Human molecular genetics*, 24: R93–R101, 2015. doi:10.1093/hmg/ddv263.

[29] Karsten Suhre, Johannes Raffler, and Gabi Kastenmüller. Biochemical insights from population studies with genetics and metabolomics. *Archives of biochemistry and biophysics*, 589:168–176, 2016. doi:10.1016/j.abb.2015.09.023.

[30] Anna Köttgen, Johannes Raffler, Peggy Sekula, and Gabi Kastenmüller. Genome-Wide Association Studies of Metabolite Concentrations (mGWAS): Relevance for Nephrology. *Seminars in nephrology*, 38:151–174, 2018. doi:10.1016/j.semnephrol.2018.01.009.

[31] Jeremy K. Nicholson, Elaine Holmes, and Paul Elliott. The Metabolome-Wide Association Study: A New Look at Human Disease Risk Factors. *Journal of Proteome Research*, 7(9):3637–3638, 2008. doi:10.1021/pr8005099. PMID: 18707153.

[32] Ann-Kristin Petersen, Jan Krumsiek, Brigitte Wägele, Fabian J Theis, H-Erich Wichmann, Christian Gieger, and Karsten Suhre. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC bioinformatics*, 13:120, 2012. doi:10.1186/1471-2105-13-120.

[33] Marc Chadeau-Hyam, Timothy M D Ebbels, Ian J Brown, Queenie Chan, Jeremiah Stamler, Chiang Ching Huang, Martha L Daviglus, Hirotsugu Ueshima, Liancheng Zhao, Elaine Holmes, Jeremy K Nicholson, Paul Elliott, and Maria De Iorio. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *Journal of proteome research*, 9: 4620–4627, 2010. doi:10.1021/pr1003449.

[34] Karsten Suhre, So-Youn Shin, Ann-Kristin Petersen, Robert P Mohney, David Meredith, Brigitte Wägele, Elisabeth Altmaier, CARDIoGRAM, Panos Deloukas, Jeanette Erdmann, Elin Grundberg, Christopher J Hammond, Martin Hrabé de Angelis, Gabi Kastenmüller, Anna Köttgen, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477: 54–60, 2011. doi:10.1038/nature10354.

[35] Matthias Arnold, Johannes Raffler, Arne Pfeufer, Karsten Suhre, and Gabi Kastenmüller. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics (Oxford, England)*, 31:1334–1336, 2015. doi:10.1093/bioinformatics/btu779.

[36] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012. doi:10.1038/nature11247.

[37] Idan Gabdank, Esther T Chan, Jean M Davidson, Jason A Hilton, Carrie A Davis, Ulugbek K Baymuradov, Aditi Narayanan, Kathrina C Onate, Keenan Graham, Stuart R Miyasato, Timothy R Dreszer, J Seth Strattan, Otto Jolanki, Forrest Y Tanaka, Benjamin C Hitz, et al. Prevention of data duplication for high throughput sequencing repositories. *Database : the journal of biological databases and curation*, 2018, 2018. doi:10.1093/database/bay008.

[38] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45:D353–D361, 2017. doi:10.1093/nar/gkw1092.

[39] Shujiro Okuda, Takuji Yamada, Masami Hamajima, Masumi Itoh, Toshiaki Katayama, Peer Bork, Susumu Goto, and Minoru Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic acids research*, 36: W423–W426, 2008. doi:10.1093/nar/gkn282.

[40] Werner Römisch-Margl, Cornelia Prehn, Ralf Bogumil, Cornelia Röhring, Karsten Suhre, and Jerzy Adamski. Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. *Metabolomics*, 8(1):133–142, 2012. doi:10.1007/s11306-011-0293-4.

[41] Stephen J Bruce, Isabelle Tavazzi, Véronique Parisod, Serge Rezzi, Sunil Kochhar, and Philippe A Guy. Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Analytical chemistry*, 81:3285–3296, 2009. doi:10.1021/ac8024569.

[42] Dajana Vuckovic. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Analytical and bioanalytical chemistry*, 403:1523–1548, 2012. doi:10.1007/s00216-012-6039-y.

[43] Mark Haid, Caroline Muschet, Simone Wahl, Werner Römisch-Margl, Cornelia Prehn, Gabriele Möller, and Jerzy Adamski. Long-Term Stability of Human Plasma Metabolites during Storage at -80 °C. *Journal of proteome research*, 17:203–211, 2018. doi:10.1021/acs.jproteome.7b00518.

[44] Stefan Kg Grebe and Ravinder J Singh. LC-MS/MS in the Clinical Laboratory – Where to From Here? *The Clinical biochemist. Reviews*, 32:5–31, 2011.

[45] John W Honour. Benchtop mass spectrometry in clinical biochemistry. *Annals of clinical biochemistry*, 40:628–638, 2003. doi:10.1258/000456303770367216.

[46] W J Griffiths, A P Jonsson, S Liu, D K Rai, and Y Wang. Electrospray and tandem mass spectrometry in biochemistry. *The Biochemical journal*, 355: 545–561, 2001.

[47] Matthias Wilm. Principles of electrospray ionization. *Molecular & cellular proteomics : MCP*, 10:M111.009407, 2011. doi:10.1074/mcp.M111.009407.

[48] Matthias S Wilm and Matthias Mann. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes*, 136(2):167 – 180, 1994. doi:10.1016/0168-1176(94)04024-9.

[49] G J Van Berkel. Electrolytic deposition of metals on to the high-voltage contact in an electrospray emitter: implications for gas-phase ion formation. *Journal of mass spectrometry : JMS*, 35:773–783, 2000. doi:10.1002/1096-9888(200007)35:7<773::AID-JMS4>3.0.CO;2-6.

[50] Andreas Wieser, Lukas Schneider, Jette Jung, and Sören Schubert. MALDI-TOF MS in microbiological diagnostics-identification of microorganisms and beyond (mini review). *Applied microbiology and biotechnology*, 93:965–974, 2012. doi:10.1007/s00253-011-3783-4.

[51] Michael Guilhaus. Special feature: Tutorial. Principles and instrumentation in time-of-flight mass spectrometry. Physical and instrumental concepts. *Journal of Mass Spectrometry*, 30(11):1519–1532, 1995. doi:10.1002/jms.1190301102.

[52] Jae C Schwartz, Michael W Senko, and John E P Syka. A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry*, 13:659–669, 2002. doi:10.1016/S1044-0305(02)00384-7.

[53] Loboda, Krutchinsky, Bromirski, Ens, and Standing. A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser

desorption/ionization source: design and performance. *Rapid communications in mass spectrometry : RCM*, 14:1047–1057, 2000. doi:10.1002/1097-0231(20000630)14:12<1047::AID-RCM990>3.0.CO;2-E.

[54] Raymond E March. An Introduction to Quadrupole Ion Trap Mass Spectrometry. *Journal of Mass Spectrometry*, 32(4):351–369, 1997. doi:10.1002/(SICI)1096-9888(199704)32:4<351::AID-JMS512>3.0.CO;2-Y.

[55] Mark Hardman and Alexander A Makarov. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Analytical chemistry*, 75:1699–1705, 2003. doi:10.1021/ac0258047.

[56] Michaela Scigelova and Alexander Makarov. Orbitrap mass analyzer–overview and applications in proteomics. *Proteomics*, 6 Suppl 2:16–21, 2006. doi:10.1002/pmic.200600528.

[57] Gary L Glish and Richard W Vachet. The basics of mass spectrometry in the twenty-first century. *Nature reviews. Drug discovery*, 2:140–150, 2003. doi:10.1038/nrd1011.

[58] Erin J Finehout and Kelvin H Lee. An introduction to mass spectrometry applications in biological research. *Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology*, 32:93–100, 2004. doi:10.1002/bmb.2004.494032020331.

[59] Ángel García-Bermejo, Manuela Ábalos, Jordi Sauló, Esteban Abad, María José González, and Belén Gómara. Triple quadrupole tandem mass spectrometry: A real alternative to high resolution magnetic sector instrument for the analysis of polychlorinated dibenzo-p-dioxins, furans and dioxin-like polychlorinated biphenyls. *Analytica chimica acta*, 889:156–165, 2015. doi:10.1016/j.aca.2015.07.039.

[60] Bo Wen, Li Ma, Sidney D Nelson, and Mingshe Zhu. High-throughput screening and characterization of reactive metabolites using polarity switching of hybrid triple quadrupole linear ion trap mass spectrometry. *Analytical chemistry*, 80: 1788–1799, 2008. doi:10.1021/ac702232r.

[61] C Borchers, C E Parker, L J Deterding, and K B Tomer. Preliminary comparison of precursor scans and liquid chromatography-tandem mass spectrometry

on a hybrid quadrupole time-of-flight mass spectrometer. *Journal of Chromatography A*, 854:119–130, 1999. doi:10.1016/S0021-9673(99)00479-3.

[62] Mira Petrović, Maria Dolores Hernando, M Silvia Díaz-Cruz, and Damià Barceló. Liquid chromatography–tandem mass spectrometry for the analysis of pharmaceutical residues in environmental samples: a review. *Journal of Chromatography A*, 1067:1–14, 2005. doi:10.1016/j.chroma.2004.10.110.

[63] M Jemal. High-throughput quantitative bioanalysis by LC/MS/MS. *Biomedical chromatography : BMC*, 14:422–429, 2000. doi:10.1002/1099-0801(200010)14:6<422::AID-BMC25>3.0.CO;2-I.

[64] William J Griffiths, Therese Koal, Yuqin Wang, Matthias Kohl, David P Enot, and Hans-Peter Deigner. Targeted metabolomics for biomarker discovery. *Angewandte Chemie (International ed. in English)*, 49:5426–5445, 2010. doi:10.1002/anie.200905579.

[65] Michaela Breier, Simone Wahl, Cornelia Prehn, Marina Fugmann, Uta Ferrari, Michaela Weise, Friederike Banning, Jochen Seissler, Harald Grallert, Jerzy Adamski, and Andreas Lechner. Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. *PloS one*, 9:e89728, 2014. doi:10.1371/journal.pone.0089728.

[66] Juan Manuel Cevallos-Cevallos and José Ignacio Reyes-De-Corcuera. Metabolomics in food science. *Advances in food and nutrition research*, 67:1–24, 2012. doi:10.1016/B978-0-12-394598-3.00001-0.

[67] Zhentian Lei, David V Huhman, and Lloyd W Sumner. Mass spectrometry strategies in metabolomics. *The Journal of biological chemistry*, 286:25435–25442, 2011. doi:10.1074/jbc.R111.238691.

[68] Tobias Fuhrer and Nicola Zamboni. High-throughput discovery metabolomics. *Current opinion in biotechnology*, 31:73–78, 2015. doi:10.1016/j.copbio.2014.08.006.

[69] Gary J Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews. Molecular cell biology*, 13:263–269, 2012. doi:10.1038/nrm3314.

[70] Katja Dettmer, Pavel A Aronov, and Bruce D Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26:51–78, 2007. doi:10.1002/mas.20108.

[71] David S Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, Daniel Johnson, Carin Li, Naama Karu, Zinat Sayeeda, Elvis Lo, Nazanin Assempour, Mark Berjanskii, Sandeep Singhal, et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic acids research*, 46:D608–D617, 2018. doi:10.1093/nar/gkx1089.

[72] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem Substance and Compound databases. *Nucleic acids research*, 44:D1202–D1213, 2016. doi:10.1093/nar/gkv951.

[73] Antony Williams and Valery Tkachenko. The Royal Society of Chemistry and the delivery of chemistry data repositories for the community. *Journal of computer-aided molecular design*, 28:1023–1030, 2014. doi:10.1007/s10822-014-9784-5.

[74] K. Murray Kermit, K. Boyd Robert, N. Eberlin Marcos, Langley G. John, Li Liang, and Naito Yasuhide. Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013), 2013.

[75] Tobias Kind and Oliver Fiehn. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*, 7:234, 2006. doi:10.1186/1471-2105-7-234.

[76] Anthony W T Bristow and Kenneth S Webb. Intercomparison study on accurate mass measurement of small molecules in mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 14:1086–1098, 2003. doi:10.1016/S1044-0305(03)00403-3.

[77] A Gareth Brenton and A Ruth Godfrey. Accurate mass measurement: terminology and treatment of data. *Journal of the American Society for Mass Spectrometry*, 21:1821–1835, 2010. doi:10.1016/j.jasms.2010.06.006.

[78] Alexandros P Siskos, Pooja Jain, Werner Römisch-Margl, Mark Bennett, David Achaintre, Yasmin Asad, Luke Marney, Larissa Richardson, Albert Koulman,

Julian L Griffin, Florence Raynaud, Augustin Scalbert, Jerzy Adamski, Cornelia Prehn, and Hector C Keun. Interlaboratory Reproducibility of a Targeted Metabolomics Platform for Analysis of Human Serum and Plasma. *Analytical chemistry*, 89:656–665, 2017. doi:10.1021/acs.analchem.6b02930.

[79] André B Canelas, Nicola Harrison, Alessandro Fazio, Jie Zhang, Juha-Pekka Pitkänen, Joost van den Brink, Barbara M Bakker, Lara Bogner, Jildau Bouwman, Juan I Castrillo, Ayca Cankorur, Pramote Chumnanpuen, Pascale Daran-Lapujade, Duygu Dikicioglu, Karen van Eunen, et al. Integrated multilaboratory systems biology reveals differences in protein metabolism between two reference yeast strains. *Nature communications*, 1:145, 2010. doi:10.1038/ncomms1150.

[80] J T Wu. Screening for inborn errors of amino acid metabolism. *Annals of clinical and laboratory science*, 21:123–142, 1991.

[81] Martin Lindner, Andreas Schulze, Siegfried Zabransky, Kurt Engelhorn, and Georg Friedrich Hoffmann. Früherkennung von Stoffwechselerkrankungen – Neue Entwicklungen im Neugeborenenscreening. *Medizinische Genetik*, 13(4): 342–347, 2001.

[82] Jens Nielsen. Systems Biology of Metabolism. *Annual review of biochemistry*, 86:245–275, 2017. doi:10.1146/annurev-biochem-061516-044757.

[83] S E Stein and D R Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5:859–866, 1994. doi:10.1016/1044-0305(94)87009-8.

[84] Kerstin Scheubert, Franziska Hufsky, and Sebastian Böcker. Computational mass spectrometry for small molecules. *Journal of cheminformatics*, 5:12, 2013. doi:10.1186/1758-2946-5-12.

[85] Ralf Tautenhahn, Kevin Cho, Winnie Uritboonthai, Zhengjiang Zhu, Gary J Patti, and Gary Siuzdak. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nature biotechnology*, 30:826–828, 2012. doi:10.1038/nbt.2348.

[86] Rónán Daly, Simon Rogers, Joe Wandy, Andris Jankevics, Karl E V Burgess, and Rainer Breitling. MetAssign: probabilistic annotation of metabolites from

LC-MS data using a Bayesian clustering approach. *Bioinformatics (Oxford, England)*, 30:2764–2771, 2014. doi:10.1093/bioinformatics/btu370.

[87] Felicity Allen, Russ Greiner, and David Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, 2015. doi:10.1007/s11306-014-0676-4.

[88] Christoph Ruttkies, Emma L Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics*, 8:3, 2016. doi:10.1186/s13321-016-0115-9.

[89] William M B Edmands, Lauren Petrick, Dinesh K Barupal, Augustin Scalbert, Mark J Wilson, Jeffrey K Wickliffe, and Stephen M Rappaport. compMS2Miner: An Automatable Metabolite Identification, Visualization, and Data-Sharing R Package for High-Resolution LC-MS Data Sets. *Analytical chemistry*, 89:3919–3928, 2017. doi:10.1021/acs.analchem.6b02394.

[90] Markus Heinonen, Huibin Shen, Nicola Zamboni, and Juho Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics (Oxford, England)*, 28:2333–2341, 2012. doi:10.1093/bioinformatics/bts437.

[91] Huibin Shen, Kai Dührkop, Sebastian Böcker, and Juho Rousu. Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics (Oxford, England)*, 30:i157–i164, 2014. doi:10.1093/bioinformatics/btu275.

[92] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America*, 112:12580–12585, 2015. doi:10.1073/pnas.1509788112.

[93] Dai Hai Nguyen, Canh Hao Nguyen, and Hiroshi Mamitsuka. SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics (Oxford, England)*, 34:i323–i332, 2018. doi:10.1093/bioinformatics/bty252.

[94] Dmitry Grapov, Kwanjeera Wanichthanarak, and Oliver Fiehn. MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. *Bioinformatics (Oxford, England)*, 31:2757–2760, 2015. doi:10.1093/bioinformatics/btv194.

[95] Jens Nielsen. Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. *Cell metabolism*, 25:572–579, 2017. doi:10.1016/j.cmet.2017.02.002.

[96] Thijs Grunhagen, Aboulfazl Shirazi-Adl, Jeremy C T Fairbank, and Jill P G Urban. Intervertebral disk nutrition: a review of factors influencing concentrations of nutrients and metabolites. *The Orthopedic clinics of North America*, 42:465–77, vii, 2011. doi:10.1016/j.ocl.2011.07.010.

[97] Nikolaos Psychogios, David D Hau, Jun Peng, An Chi Guo, Rupasri Mandal, Souhaila Bouatra, Igor Sinelnikov, Ramanarayan Krishnamurthy, Roman Eisner, Bijaya Gautam, Nelson Young, Jianguo Xia, Craig Knox, Edison Dong, Paul Huang, et al. The human serum metabolome. *PloS one*, 6:e16957, 2011. doi:10.1371/journal.pone.0016957.

[98] Lindsay M Edwards. Metabolic systems biology: a brief primer. *The Journal of physiology*, 595:2849–2855, 2017. doi:10.1113/JP272275.

[99] Bensu Karahalil. Overview of Systems Biology and Omics Technologies. *Current medicinal chemistry*, 23:4221–4230, 2016. doi:10.2174/0929867323666160926150617.

[100] Peter Kohl and Denis Noble. Systems biology and the virtual physiological human. *Molecular systems biology*, 5:292, 2009. doi:10.1038/msb.2009.51.

[101] Francis Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970. doi:10.1038/227561a0.

[102] Wolfgang Epstein and Jonathan R Beckwith. Regulation of gene expression. *Annual Review of Biochemistry*, 37(1):411–436, 1968. doi:10.1146/annurev.bi.37.070168.002211.

[103] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33 Suppl:245–254, 2003. doi:10.1038/ng1089.

[104] Quan Zhong, Samuel J Pevzner, Tong Hao, Yang Wang, Roberto Mosca, Jörg Menche, Mikko Taipale, Murat Taşan, Changyu Fan, Xinping Yang, Patrick Haley, Ryan R Murray, Flora Mer, Fana Gebreab, Stanley Tam, et al. An interspecies protein-protein interaction network across vast evolutionary distance. *Molecular systems biology*, 12:865, 2016. doi:10.15252/msb.20156484.

[105] Daniel R Hyduke, Nathan E Lewis, and Bernhard Ø Palsson. Analysis of omics data with genome-scale models of metabolism. *Molecular bioSystems*, 9:167–174, 2013. doi:10.1039/c2mb25453k.

[106] Albert-László Barabási. Scale-free networks: a decade and beyond. *Science (New York, N.Y.)*, 325:412–413, 2009. doi:10.1126/science.1173299.

[107] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12:56–68, 2011. doi:10.1038/nrg2918.

[108] István A Kovács and Albert-László Barabási. Network science: Destruction perfected. *Nature*, 524:38–39, 2015. doi:10.1038/524038a.

[109] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5:101–113, 2004. doi:10.1038/nrg1272.

[110] Ines Thiele, Neil Swainston, Ronan M T Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, Stefan G Thorleifsson, Rasmus Agren, Christian Bölling, Sergio Bordel, Arvind K Chavali, et al. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31:419–425, 2013. doi:10.1038/nbt.2488.

[111] Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, Andreas Prlić, Anand Sastry, Anna D Danielsdottir, Almut Heinken, Alberto Noronha, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36:272–281, 2018. doi:10.1038/nbt.4072.

[112] Antonio Rosato, Leonardo Tenori, Marta Cascante, Pedro Ramon De Atauri Carulla, Vitor A P Martins Dos Santos, and Edoardo Saccenti. From

correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics : Official journal of the Metabolomic Society*, 14:37, 2018. doi:10.1007/s11306-018-1335-y.

[113] Khuram Shahzad and Juan J. Loor. Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Current Genomics*, 13(PMC3401895):379–394, 2012. doi:10.2174/138920212801619269.

[114] Ron Caspi, Richard Billington, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E Midford, Quang Ong, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes. *Nucleic acids research*, 46:D633–D639, 2018. doi:10.1093/nar/gkx935.

[115] Pedro Romero, Jonathan Wagg, Michelle L Green, Dale Kaiser, Markus Krummenacker, and Peter D Karp. Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, 6:R2, 2004. doi:10.1186/gb-2004-6-1-r2.

[116] F Hoffmann-La Roche Ltd Corporate Donations and Philanthropy, 4070 Basel, Switzerland. Roche Biochemical Pathways, 4th Edition, 2014. URL `http://biochemical-pathways.com`.

[117] Sandra Placzek, Ida Schomburg, Antje Chang, Lisa Jeske, Marcus Ulbrich, Jana Tillack, and Dietmar Schomburg. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic acids research*, 45:D380–D388, 2017. doi:10.1093/nar/gkw952.

[118] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5:21, 2011. doi:10.1186/1752-0509-5-21.

[119] Jeremy K Nicholson and John C Lindon. Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056, 2008. doi:10.1038/4551054a.

[120] Joseph Loscalzo and Albert-Laszlo Barabasi. Systems biology and the future of medicine. *Wiley interdisciplinary reviews. Systems biology and medicine*, 3: 619–627, 2011. doi:10.1002/wsbm.144.

[121] Mika Gustafsson, Colm E Nestor, Huan Zhang, Albert-László Barabási, Sergio Baranzini, Sören Brunak, Kian Fan Chung, Howard J Federoff, Anne-Claude Gavin, Richard R Meehan, Paola Picotti, Miguel Àngel Pujana, Nikolaus Rajewsky, Kenneth Gc Smith, Peter J Sterk, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome medicine*, 6: 82, 2014. doi:10.1186/s13073-014-0082-6.

[122] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science (New York, N.Y.)*, 347:1257601, 2015. doi:10.1126/science.1257601.

[123] J Larry Jameson and Dan L Longo. Precision medicine–personalized, problematic, and promising. *The New England journal of medicine*, 372:2229–2234, 2015. doi:10.1056/NEJMsb1503104.

[124] Richard D Beger, Warwick Dunn, Michael A Schmidt, Steven S Gross, Jennifer A Kirwan, Marta Cascante, Lorraine Brennan, David S Wishart, Matej Oresic, Thomas Hankemeier, David I Broadhurst, Andrew N Lane, Karsten Suhre, Gabi Kastenmüller, Susan J Sumner, et al. Metabolomics enables precision medicine: "A White Paper, Community Perspective". *Metabolomics : Official journal of the Metabolomic Society*, 12:149, 2016. doi:10.1007/s11306-016-1094-6.

[125] Peiyuan Yin and Guowang Xu. Current state-of-the-art of nontargeted metabolomics based on liquid chromatography-mass spectrometry with special emphasis in clinical applications. *Journal of Chromatography A*, 1374:1–13, 2014. doi:10.1016/j.chroma.2014.11.050.

[126] Mark R Viant, Irwin J Kurland, Martin R Jones, and Warwick B Dunn. How close are we to complete annotation of metabolomes? *Current opinion in chemical biology*, 36:64–69, 2017. doi:10.1016/j.cbpa.2017.01.001.

[127] Darren J Creek, Warwick B Dunn, Oliver Fiehn, Julian L Griffin, Robert D Hall, Zhentian Lei, Robert Mistrik, Steffen Neumann, Emma L Schymanski, Lloyd W Sumner, Robert Trengove, and Jean-Luc Wolfender. Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics*, 10(3):350–353, 2014. doi:10.1007/s11306-014-0656-8.

[128] Lloyd W Sumner, Zhentian Lei, Basil J Nikolau, Kazuki Saito, Ute Roessner, and Robert Trengove. Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics*, 10(6):1047–1049, 2014. doi:10.1007/s11306-014-0739-6.

[129] Gerrit van Meer, Dennis R Voelker, and Gerald W Feigenson. Membrane lipids: where they are and how they behave. *Nature reviews. Molecular cell biology*, 9: 112–124, 2008. doi:10.1038/nrm2330.

[130] Katharina Leidl, Gerhard Liebisch, Dorothea Richter, and Gerd Schmitz. Mass spectrometric analysis of lipid species of human circulating blood cells. *Biochimica et biophysica acta*, 1781:655–664, 2008. doi:10.1016/j.bbalip.2008.07.008.

[131] Laura K Cole, Jean E Vance, and Dennis E Vance. Phosphatidylcholine biosynthesis and lipoprotein metabolism. *Biochimica et biophysica acta*, 1821: 754–761, 2012. doi:10.1016/j.bbalip.2011.09.009.

[132] Jelske N van der Veen, John P Kennelly, Sereana Wan, Jean E Vance, Dennis E Vance, and René L Jacobs. The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. *Biochimica et biophysica acta*, 1859:1558–1572, 2017. doi:10.1016/j.bbamem.2017.04.006.

[133] Peiyuan Yin, Patamu Mohemaiti, Jing Chen, Xinjie Zhao, Xin Lu, Adilijiang Yimiti, Halmurat Upur, and Guowang Xu. Serum metabolic profiling of abnormal savda by liquid chromatography/mass spectrometry. *Journal of Chromatography B, Analytical technologies in the biomedical and life sciences*, 871: 322–327, 2008. doi:10.1016/j.jchromb.2008.05.043.

[134] J Peterson, B E Bihain, G Bengtsson-Olivecrona, R J Deckelbaum, Y A Carpentier, and T Olivecrona. Fatty acid control of lipoprotein lipase: a link between energy metabolism and lipid transport. *Proceedings of the National Academy of Sciences of the United States of America*, 87:909–913, 1990. doi:10.1073/pnas.87.3.909.

[135] Artemis P Simopoulos. The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases. *Experimental biology and medicine (Maywood, N.J.)*, 233:674–688, 2008. doi:10.3181/0711-MR-311.

[136] Masato Abe, Yoshiki Sawada, Shinpei Uno, Shuhei Chigasaki, Masahide Oku, Yasuyoshi Sakai, and Hideto Miyoshi. Role of Acyl Chain Composition of Phosphatidylcholine in Tafazzin-Mediated Remodeling of Cardiolipin in Liposomes. *Biochemistry*, 56:6268–6280, 2017. doi:10.1021/acs.biochem.7b00941.

[137] Masahiro Nagahama, Akiko Otsuka, Masataka Oda, Rajesh K Singh, Zyta M Ziora, Hiroshi Imagawa, Mugio Nishizawa, and Jun Sakurai. Effect of unsaturated bonds in the sn-2 acyl chain of phosphatidylcholine on the membrane-damaging action of Clostridium perfringens alpha-toxin toward liposomes. *Biochimica et biophysica acta*, 1768:2940–2945, 2007. doi:10.1016/j.bbamem.2007.08.016.

[138] Benjamin Jenkins, James A West, and Albert Koulman. A Review of Odd-Chain Fatty Acid Metabolism and the Role of Pentadecanoic Acid (C15:0) and Heptadecanoic Acid (C17:0) in Health and Disease. *Molecules (Basel, Switzerland)*, 20:2425–2444, 2015. doi:10.3390/molecules20022425.

[139] L Marai and A Kuksis. Molecular species of lecithins from erythrocytes and plasma of man. *Journal of Lipid Research*, 10:141–152, 1969.

[140] Karsten Suhre, Werner Römisch-Margl, Martin Hrabé de Angelis, Jerzy Adamski, Gerd Luippold, and Robert Augustin. Identification of a potential biomarker for FABP4 inhibition: the power of lipidomics in preclinical drug testing. *Journal of biomolecular screening*, 16:467–475, 2011. doi:10.1177/1087057111402200.

[141] Anna Floegel, Norbert Stefan, Zhonghao Yu, Kristin Mühlenbruch, Dagmar Drogan, Hans-Georg Joost, Andreas Fritsche, Hans-Ulrich Häring, Martin Hrabe de Angelis, Annette Peters, Michael Roden, Cornelia Prehn, Rui Wang-Sattler, Thomas Illig, Matthias B Schulze, et al. Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes*, 62:639–648, 2013. doi:10.2337/db12-0495.

[142] Luke Whiley, Arundhuti Sen, James Heaton, Petroula Proitsi, Diego García-Gómez, Rufina Leung, Norman Smith, Madhav Thambisetty, Iwona Kloszewska, Patrizia Mecocci, Hilkka Soininen, Magda Tsolaki, Bruno Vellas, Simon Lovestone, Cristina Legido-Quigley, et al. Evidence of altered phosphatidylcholine metabolism in Alzheimer's disease. *Neurobiology of aging*, 35:271–278, 2014. doi:10.1016/j.neurobiolaging.2013.08.001.

[143] Jon B Toledo, Matthias Arnold, Gabi Kastenmüller, Rui Chang, Rebecca A Baillie, Xianlin Han, Madhav Thambisetty, Jessica D Tenenbaum, Karsten Suhre, J Will Thompson, Lisa St John-Williams, Siamak MahmoudianDehkordi, Daniel M Rotroff, John R Jack, Alison Motsinger-Reif, et al. Metabolic network failures in Alzheimer's disease: A biochemical road map. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 13:965–984, 2017. doi:10.1016/j.jalz.2017.01.020.

[144] Harmen H M Draisma, René Pool, Michael Kobl, Rick Jansen, Ann-Kristin Petersen, Anika A M Vaarhorst, Idil Yet, Toomas Haller, Ayse Demirkan, Tõnu Esko, Gu Zhu, Stefan Böhringer, Marian Beekman, Jan Bert van Klinken, Werner Römisch-Margl, et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nature communications*, 6:7208, 2015. doi:10.1038/ncomms8208.

[145] Sergio Lario, Maria José Ramírez-Lázaro, Daniel Sanjuan-Herráez, Anna Brunet-Vega, Carles Pericay, Lourdes Gombau, Félix Junquera, Guillermo Quintás, and Xavier Calvet. Plasma sample based analysis of gastric cancer progression using targeted metabolomics. *Scientific reports*, 7:17774, 2017. doi:10.1038/s41598-017-17921-x.

[146] Erwan Werner, Jean-François Heilier, Céline Ducruix, Eric Ezan, Christophe Junot, and Jean-Claude Tabet. Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. *Journal of Chromatography B, Analytical technologies in the biomedical and life sciences*, 871:143–163, 2008. doi:10.1016/j.jchromb.2008.07.004.

[147] Jan Krumsiek, Karsten Suhre, Anne M Evans, Matthew W Mitchell, Robert P Mohney, Michael V Milburn, Brigitte Wägele, Werner Römisch-Margl, Thomas Illig, Jerzy Adamski, Christian Gieger, Fabian J Theis, and Gabi Kastenmüller. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS genetics*, 8:e1003005, 2012. doi:10.1371/journal.pgen.1003005.

[148] Rainer Breitling, Shawn Ritchie, Dayan Goodenowe, Mhairi L Stewart, and Michael P Barrett. Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics : Official journal of the Metabolomic Society*, 2:155–164, 2006. doi:10.1007/s11306-006-0029-z.

[149] Jan D Quell, Werner Römisch-Margl, Mark Haid, Jan Krumsiek, Thomas Skurk, Jerzy Adamski, Hans Hauner, Robert Mohney, Hannelore Daniel, Karsten Suhre, and Gabi Kastenmüller. A platform comparison to characterize compositions of phosphatidylcholines measured as lipid sums in human plasma. *Manuscript for concurrent submission to the Journal of Lipid Research.*

[150] Jan D Quell, Werner Römisch-Margl, Marco Colombo, Jan Krumsiek, Anne M Evans, Robert Mohney, Veikko Salomaa, Ulf de Faire, Leif C Groop, Felix Agakov, Helen C Looker, Paul McKeigue, Helen M Colhoun, and Gabi Kastenmüller. Automated pathway and reaction prediction facilitates in silico identification of unknown metabolites in human cohort studies. *Journal of Chromatography B, Analytical technologies in the biomedical and life sciences*, 1071: 58–67, 2017. doi:10.1016/j.jchromb.2017.04.002.

[151] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, R version 3.4.0 edition, 2017. URL `http://www.R-project.org/`.

[152] E Ullah, M Shahzad, R Rawi, M Dehbi, K Suhre, M Selim, and H Bensmail. Integrative ^{1}H-NMR-based Metabolomic Profiling to Identify Type-2 Diabetes Biomarkers: An Application to a Population of Qatar. *Metabolomics:Open Access*, 5(1):1–8, 2015. doi:10.4172/2153-0769.1000136.

[153] Baljit K Ubhi, Alex Conner, Eva Duchoslav, Annie Evans, Richard Robinson, Leo Wang, Paul RS Baker, and Steve Watkins. A Novel Lipid Screening Platform that Provides a Complete Solution for Lipidomics Research. *AB SCIEX Technical Application Note, RUO-MKT-02-2871B*, 2016.

[154] Baljit K Ubhi. Novel Chemical Standards Kits Enable Facile Lipid Quantitation. *AB SCIEX Technical Application Note, RUO-MKT-02-3879-A*, 2016.

[155] Paul R S Baker, Aaron M Armando, J Larry Campbell, Oswald Quehenberger, and Edward A Dennis. Three-dimensional enhanced lipidomics analysis combining UPLC, differential ion mobility spectrometry, and mass spectrometric separation strategies. *Journal of Lipid Research*, 55:2432–2442, 2014. doi:10.1194/jlr.D051581.

[156] Tuulia P I Lintonen, Paul R S Baker, Matti Suoniemi, Baljit K Ubhi, Kaisa M Koistinen, Eva Duchoslav, J Larry Campbell, and Kim Ekroos. Differential mo-

bility spectrometry-driven shotgun lipidomics. *Analytical chemistry*, 86:9662–9669, 2014. doi:10.1021/ac5021744.

[157] E G Bligh and W J Dyer. A rapid method of total lipid extraction and purification. *Canadian journal of biochemistry and physiology*, 37:911–917, 1959. doi:10.1139/o59-099.

[158] Yves LeBlanc, Doina Caraiman, Mauro Aiello, and Hesham Ghobarah. SelexION Technology: A New Solution to Selectivity Challenges in Quantitative Bioanalysis. *AB SCIEX Technical Note, RUO-MKT-02-3251-A*, 2015.

[159] Paul RS Baker, J Larry Campbell, Eva Duchoslav, and Christie Hunter. Differential Mobility Separation for Improving Lipidomic Analysis by Mass Spectrometry. *AB SCIEX Technical Application Note, RUO-MKT-02-4802-A*, 2017.

[160] Baljit K Ubhi. Direct Infusion-Tandem Mass Spectrometry (DI-MS/MS) Analysis of Complex Lipids in Human Plasma and Serum Using the Lipidyzer^{TM} Platform. *Methods in molecular biology (Clifton, N.J.)*, 1730:227–236, 2018. doi:10.1007/978-1-4939-7592-1_15.

[161] Andras Franko, Dietrich Merkel, Marketa Kovarova, Miriam Hoene, Benjamin A Jaghutriz, Martin Heni, Alfred Königsrainer, Cyrus Papan, Stefan Lehr, Hans-Ulrich Häring, and Andreas Peter. Dissociation of Fatty Liver and Insulin Resistance in I148M PNPLA3 Carriers: Differences in Diacylglycerol (DAG) FA18:1 Lipid Species as a Possible Explanation. *Nutrients*, 10, 2018. doi:10.3390/nu10091314.

[162] Gerhard Liebisch, Juan Antonio Vizcaíno, Harald Köfeler, Martin Trötzmüller, William J Griffiths, Gerd Schmitz, Friedrich Spener, and Michael J O Wakelam. Shorthand notation for lipid structures derived from mass spectrometry. *Journal of Lipid Research*, 54:1523–1530, 2013. doi:10.1194/jlr.M033506.

[163] Anne M Evans, Corey D DeHaven, Tom Barrett, Matt Mitchell, and Eric Milgram. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical chemistry*, 81:6656–6667, 2009. doi:10.1021/ac901536h.

[164] Kurt J Boudonck, Matthew W Mitchell, Jacob Wulff, and John A Ryals. Characterization of the biochemical variability of bovine milk using metabolomics. *Metabolomics*, 5(4):375–386, 2009. doi:10.1007/s11306-009-0160-8.

[165] AM Evans, BR Bridgewater, Q Liu, MW Mitchell, RJ Robinson, H Dai, SJ Stewart, CD DeHaven, and LAD Miller. High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High- Throughput Profiling Metabolomics. *Metabolomics*, 4(132):1–7, 2014. doi:10.4172/2153-0769.1000132.

[166] Roman A Zubarev and Alexander Makarov. Orbitrap mass spectrometry. *Analytical chemistry*, 85:5288–5296, 2013. doi:10.1021/ac4001223.

[167] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, et al. An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46:543–550, 2014. doi:10.1038/ng.2982.

[168] Scott Saobing Chen and Ramesh A Gopinath. Gaussianization. In T K Leen, T G Dietterich, and V Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 423–429. MIT Press, 2001. URL `http://papers.nips.cc/paper/1856-gaussianization.pdf`.

[169] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, pages 1–68, 2010. URL `https://www.jstatsoft.org/article/view/v045i03`.

[170] Juliane Schaefer, Rainer Opgen-Rhein, and Korbinian Strimmer. GeneNet: modeling and inferring gene networks. 1(8), 2013. URL `https://cran.r-project.org/web/packages/GeneNet/`.

[171] J van Helden, L Wernisch, D Gilbert, and S J Wodak. Graph-based analysis of metabolic networks. *Ernst Schering Research Foundation workshop*, pages 245–274, 2002. doi:10.1007/978-3-662-04747-7_12.

[172] Vincent Lacroix, Ludovic Cottret, Patricia Thébault, and Marie-France Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, 5:594–617, 2008. doi:10.1109/TCBB.2008.79.

[173] Clément Frainay and Fabien Jourdan. Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Briefings in bioinformatics*, 18:43–56, 2017. doi:10.1093/bib/bbv115.

[174] Pascal Schlosser, ..., and Anna Köttgen. Genome-wide Association Studies of Urine Metabolite Concentrations among CKD Patients Identifies Multiple Loci Related to Metabolite Excretion and Detoxification. *Manuscript in preparation.*

[175] Manish Sud, Eoin Fahy, Dawn Cotter, Alex Brown, Edward A Dennis, Christopher K Glass, Alfred H Merrill, Robert C Murphy, Christian R H Raetz, David W Russell, and Shankar Subramaniam. LMSD: LIPID MAPS structure database. *Nucleic acids research*, 35:D527–D532, 2007. doi:10.1093/nar/gkl838.

[176] PubChem Data Specification. *PubChem Substructure Fingerprint*, 2009. URL `ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_sdtags.pdf`.

[177] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:20, 2015. doi:10.1186/s13321-015-0069-3.

[178] Yan Wang, Tyler W H Backman, Kevin Horan, and Thomas Girke. fmcsR: mismatch tolerant maximum common substructure searching in R. *Bioinformatics (Oxford, England)*, 29:2792–2794, 2013. doi:10.1093/bioinformatics/btt475.

[179] Yiqun Cao, Anna Charisi, Li-Chang Cheng, Tao Jiang, and Thomas Girke. ChemmineR: a compound mining framework for R. *Bioinformatics (Oxford, England)*, 24:1733–1734, 2008. doi:10.1093/bioinformatics/btn307.

[180] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, et al. *caret: Classification and Regression Training*, 2017. URL `https://CRAN.R-project.org/package=caret`. R package version 6.0-78.

[181] Idil Yet, Cristina Menni, So-Youn Shin, Massimo Mangino, Nicole Soranzo, Jerzy Adamski, Karsten Suhre, Tim D Spector, Gabi Kastenmüller, and Jordana T Bell. Genetic Influences on Metabolite Levels: A

Comparison across Metabolomic Platforms. *PloS one*, 11:e0153672, 2016. doi:10.1371/journal.pone.0153672.

[182] Rui Wang-Sattler, Zhonghao Yu, Christian Herder, Ana C Messias, Anna Floegel, Ying He, Katharina Heim, Monica Campillos, Christina Holzapfel, Barbara Thorand, Harald Grallert, Tao Xu, Erik Bader, Cornelia Huth, Kirstin Mittelstrass, et al. Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular systems biology*, 8:615, 2012. doi:10.1038/msb.2012.43.

[183] Janina S Ried, So-Youn Shin, Jan Krumsiek, Thomas Illig, Fabian J Theis, Tim D Spector, Jerzy Adamski, H-Erich Wichmann, Konstantin Strauch, Nicole Soranzo, Karsten Suhre, and Christian Gieger. Novel genetic associations with serum level metabolites identified by phenotype set enrichment analyses. *Human molecular genetics*, 23:5847–5857, 2014. doi:10.1093/hmg/ddu301.

[184] Karsten Suhre, Christa Meisinger, Angela Döring, Elisabeth Altmaier, Petra Belcredi, Christian Gieger, David Chang, Michael V Milburn, Walter E Gall, Klaus M Weinberger, Hans-Werner Mewes, Martin Hrabé de Angelis, H-Erich Wichmann, Florian Kronenberg, Jerzy Adamski, et al. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PloS one*, 5:e13953, 2010. doi:10.1371/journal.pone.0013953.

[185] Michal Holčapek, Magdaléna Ovčačíková, Miroslav Lísa, Eva Cífková, and Tomáš Hájek. Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples. *Analytical and bioanalytical chemistry*, 407:5033–5043, 2015. doi:10.1007/s00216-015-8528-2.

[186] Magdaléna Ovčačíková, Miroslav Lísa, Eva Cífková, and Michal Holčapek. Retention behavior of lipids in reversed-phase ultrahigh-performance liquid chromatography-electrospray ionization mass spectrometry. *Journal of chromatography. A*, 1450:76–85, 2016. doi:10.1016/j.chroma.2016.04.082.

[187] Evelyn Rampler, Harald Schoeny, Bernd M Mitic, Yasin El Abiead, Michaela Schwaiger, and Gunda Koellensperger. Simultaneous non-polar and polar lipid analysis by on-line combination of HILIC, RP and high resolution MS. *The Analyst*, 143:1250–1258, 2018. doi:10.1039/c7an01984j.

[188] Jeremy P Koelmel, Nicholas M Kroeger, Candice Z Ulmer, John A Bowden, Rainey E Patterson, Jason A Cochran, Christopher W W Beecher, Timothy J Garrett, and Richard A Yost. LipidMatch: an automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. *BMC bioinformatics*, 18:331, 2017. doi:10.1186/s12859-017-1744-3.

[189] Jürgen Hartler, Alexander Triebl, Andreas Ziegl, Martin Trötzmüller, Gerald N Rechberger, Oana A Zeleznik, Kathrin A Zierler, Federico Torta, Amaury Cazenave-Gassiot, Markus R Wenk, Alexander Fauland, Craig E Wheelock, Aaron M Armando, Oswald Quehenberger, Qifeng Zhang, et al. Deciphering lipid structures based on platform-independent decision rules. *Nature methods*, 14:1171–1174, 2017. doi:10.1038/nmeth.4470.

[190] Michael A Kochen, Matthew C Chambers, Jay D Holman, Alexey I Nesvizhskii, Susan T Weintraub, John T Belisle, M Nurul Islam, Johannes Griss, and David L Tabb. Greazy: Open-Source Software for Automated Phospholipid Tandem Mass Spectrometry Identification. *Analytical chemistry*, 88:5733–5741, 2016. doi:10.1021/acs.analchem.6b00021.

[191] Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature methods*, 12:523–526, 2015. doi:10.1038/nmeth.3393.

[192] Jörg Martin Büscher, Dominika Czernik, Jennifer Christina Ewald, Uwe Sauer, and Nicola Zamboni. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Analytical chemistry*, 81:2135–2143, 2009. doi:10.1021/ac8022857.

[193] Patrick Leuthold, Elke Schaeffeler, Stefan Winter, Florian Büttner, Ute Hofmann, Thomas E Mürdter, Steffen Rausch, Denise Sonntag, Judith Wahrheit, Falko Fend, Jörg Hennenlotter, Jens Bedke, Matthias Schwab, and Mathias Haag. Comprehensive Metabolomic and Lipidomic Profiling of Human Kidney Tissue: A Platform Comparison. *Journal of proteome research*, 16:933–944, 2017. doi:10.1021/acs.jproteome.6b00875.

[194] BIOCRATES Life Sciences AG. Annotation of potential isobaric and isomeric lipid species measured with the Absolute*IDQ*$^{\mathrm{TM}}$ p180 Kit (and p150

Kit). Technical report, 2018. URL `https://www.biocrates.com/images/List-of-Isobaric-and-Isomeric-Lipid-Species_v1_2018.pdf`.

[195] Chung-Ho E Lau, Alexandros P Siskos, Léa Maitre, Oliver Robinson, Toby J Athersuch, Elizabeth J Want, Jose Urquiza, Maribel Casas, Marina Vafeiadi, Theano Roumeliotaki, Rosemary R C McEachan, Rafaq Azad, Line S Haug, Helle M Meltzer, Sandra Andrusaityte, et al. Determinants of the urinary and serum metabolome in children from six European populations. *BMC medicine*, 16:202, 2018. doi:10.1186/s12916-018-1190-8.

[196] Séverine Trabado, Abdallah Al-Salameh, Vincent Croixmarie, Perrine Masson, Emmanuelle Corruble, Bruno Fève, Romain Colle, Laurent Ripoll, Bernard Walther, Claire Boursier-Neyret, Erwan Werner, Laurent Becquemont, and Philippe Chanson. The human plasma-metabolome: Reference values in 800 French healthy volunteers; impact of cholesterol, gender and age. *PloS one*, 12: e0173615, 2017. doi:10.1371/journal.pone.0173615.

[197] Therese Koal, Kristaps Klavins, Daniele Seppi, Georg Kemmler, and Christian Humpel. Sphingomyelin SM(d18:1/18:0) is significantly enhanced in cerebrospinal fluid samples dichotomized by pathological amyloid-β_{42}, tau, and phospho-tau-181 levels. *Journal of Alzheimer's disease : JAD*, 44:1193–1201, 2015. doi:10.3233/JAD-142319.

[198] Sven Zukunft, Cornelia Prehn, Cornelia Röhring, Gabriele Möller, Martin Hrabě de Angelis, Jerzy Adamski, and Janina Tokarz. High-throughput extraction and quantification method for targeted metabolomics in murine tissues. *Metabolomics : Official journal of the Metabolomic Society*, 14:18, 2018. doi:10.1007/s11306-017-1312-x.

[199] Jillian M Hagel, Rupasri Mandal, Beomsoo Han, Jun Han, Donald R Dinsmore, Christoph H Borchers, David S Wishart, and Peter J Facchini. Metabolome analysis of 20 taxonomically related benzylisoquinoline alkaloid-producing plants. *BMC plant biology*, 15:220, 2015. doi:10.1186/s12870-015-0594-2.

[200] Souhaila Bouatra, Farid Aziat, Rupasri Mandal, An Chi Guo, Michael R Wilson, Craig Knox, Trent C Bjorndahl, Ramanarayan Krishnamurthy, Fozia Saleem, Philip Liu, Zerihun T Dame, Jenna Poelzer, Jessica Huynh, Faizath S Yallou, Nick Psychogios, et al. The human urine metabolome. *PloS one*, 8: e73076, 2013. doi:10.1371/journal.pone.0073076.

[201] Silvia Turroni, Jessica Fiori, Simone Rampelli, Stephanie L Schnorr, Clarissa Consolandi, Monica Barone, Elena Biagi, Flaminia Fanelli, Marco Mezzullo, Alyssa N Crittenden, Amanda G Henry, Patrizia Brigidi, and Marco Candela. Fecal metabolome of the Hadza hunter-gatherers: a host-microbiome integrative view. *Scientific reports*, 6:32826, 2016. doi:10.1038/srep32826.

[202] T Raclot, C Holm, and D Langin. Fatty acid specificity of hormone-sensitive lipase. Implication in the selective hydrolysis of triacylglycerols. *Journal of Lipid Research*, 42:2049–2057, 2001.

[203] John A Bowden, Alan Heckert, Candice Z Ulmer, Christina M Jones, Jeremy P Koelmel, Laila Abdullah, Linda Ahonen, Yazen Alnouti, Aaron M Armando, John M Asara, Takeshi Bamba, John R Barr, Jonas Bergquist, Christoph H Borchers, Joost Brandsma, et al. Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950-Metabolites in Frozen Human Plasma. *Journal of Lipid Research*, 58:2275–2288, 2017. doi:10.1194/jlr.M079012.

[204] Oswald Quehenberger, Aaron M Armando, Alex H Brown, Stephen B Milne, David S Myers, Alfred H Merrill, Sibali Bandyopadhyay, Kristin N Jones, Samuel Kelly, Rebecca L Shaner, Cameron M Sullards, Elaine Wang, Robert C Murphy, Robert M Barkley, Thomas J Leiker, et al. Lipidomics reveals a remarkable diversity of lipids in human plasma. *Journal of Lipid Research*, 51: 3299–3305, 2010. doi:10.1194/jlr.M009449.

[205] Cornelia Then, Simone Wahl, Anna Kirchhofer, Harald Grallert, Susanne Krug, Gabi Kastenmüller, Werner Römisch-Margl, Melina Claussnitzer, Thomas Illig, Margit Heier, Christa Meisinger, Jerzy Adamski, Barbara Thorand, Cornelia Huth, Annette Peters, et al. Plasma metabolomics reveal alterations of sphingo- and glycerophospholipid levels in non-diabetic carriers of the transcription factor 7-like 2 polymorphism rs7903146. *PloS one*, 8:e78430, 2013. doi:10.1371/journal.pone.0078430.

[206] Justin Johan Jozias van der Hooft, Joe Wandy, Michael P Barrett, Karl E V Burgess, and Simon Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America*, 113:13738–13743, 2016. doi:10.1073/pnas.1608041113.

[207] Sunghwan Kim, Ryan P Rodgers, and Alan G Marshall. Truly “exact” mass: Elemental composition can be determined uniquely from molecular mass measurement at ∼0.1mDa accuracy for molecules up to ∼500Da. *International Journal of Mass Spectrometry*, 251(2):260 – 265, 2006. doi:10.1016/j.ijms.2006.02.001. ULTRA-ACCURATE MASS SPECTROMETRY AND RELATED TOPICS Dedicated to H.-J. Kluge on the occasion of his 65th birthday anniversary.

[208] David G Watson. A rough guide to metabolite identification using high resolution liquid chromatography mass spectrometry in metabolomic profiling in metazoans. *Computational and structural biotechnology journal*, 4(5):e201301005, 2013. doi:10.5936/csbj.201301005.

[209] Kate Comstock, Seema Sharma, and Caroline Ding. Ultra high resolution MS for confident metabolite structure identification. *Drug Metabolism and Pharmacokinetics*, 33(1, Supplement):S24, 2018. doi:10.1016/j.dmpk.2017.11.095.

[210] Tobias Kind and Oliver Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8:105, 2007. doi:10.1186/1471-2105-8-105.

[211] Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S Mehta, Gert Wohlgemuth, Dinesh Kumar Barupal, Megan R Showalter, Masanori Arita, et al. Identification of small molecules using accurate mass MS/MS search. *Mass spectrometry reviews*, 2017. doi:10.1002/mas.21535.

[212] Robert Höllering, Johann Gasteiger, Larissa Steinhauer, Klaus-Peter Schulz, and Achim Herwig. Simulation of organic reactions: from the degradation of chemicals to combinatorial synthesis. *Journal of chemical information and computer sciences*, 40:482–494, 2000. doi:10.1021/ci990433p.

[213] Bing Yu, Yan Zheng, Danny Alexander, Alanna C Morrison, Josef Coresh, and Eric Boerwinkle. Genetic determinants influencing human serum metabolome among African Americans. *PLoS genetics*, 10:e1004212, 2014. doi:10.1371/journal.pgen.1004212.

[214] Adam D Kennedy, Bryan M Wittmann, Anne M Evans, Luke A D Miller, Douglas R Toal, Shaun Lonergan, Sarah H Elsea, and Kirk L Pappan. Metabolomics in the Clinic: A Review of the Shared and Unique Features of

Untargeted Metabolomics for Clinical Research and Clinical Testing. *Journal of mass spectrometry : JMS*, 2018. doi:10.1002/jms.4292.

[215] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry : JMS*, 45:703–714, 2010. doi:10.1002/jms.1777.

[216] Carlos Guijas, J Rafael Montenegro-Burke, Xavier Domingo-Almenara, Amelia Palermo, Benedikt Warth, Gerrit Hermann, Gunda Koellensperger, Tao Huan, Winnie Uritboonthai, Aries E Aisporna, Dennis W Wolan, Mary E Spilker, H Paul Benton, and Gary Siuzdak. METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Analytical chemistry*, 90:3156–3164, 2018. doi:10.1021/acs.analchem.7b04424.

[217] Ivana Blaženović, Tobias Kind, Jian Ji, and Oliver Fiehn. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites*, 8, 2018. doi:10.3390/metabo8020031.

APPENDIX A

Supplements: Characterization of PC sums

Supplemental tables and figures of Chapter 3 are provided as online-supplements and on the attached CD (cf. Index of external material on page 239).

Part_I_LipidCompositions/Supplemental_Table_A.1_QualitativeComp.xlsx

Table A.1: Qualitative composition of PCs measured by Absolute*IDQ*$^{\text{TM}}$
The chain length and double bonds of 76 PCs measured on the lipid species level were systematically distributed to the *sn*-1 and *sn*-2 positions to collect 150 to 456 theoretically possible molecules of the fatty acid level. If the particular SM [$^{13}C_1$]SM $x + 4 : y$ for the PC of the lipid species level PC $x : y$ has not been measured, it was included to the list. A similar version of this table is part of my prepared manuscript as well; Quell et al. [149].

Part_I_LipidCompositions/Supplemental_Table_A.2_Ranges.xlsx

Table A.2: Qualitative composition of measured PCs and ranges of metabolite levels per platform
Metabolite levels of PCs measured by Absolute*IDQ*$^{\text{TM}}$ on the lipid species level are compared to aggregated levels of related one to four PCs measured by Lipidyzer$^{\text{TM}}$ on the fatty acid level. Metabolites marked by an asterisk were excluded because of missingness $> 75\%$. This table is part of my prepared manuscript as well; Quell et al. [149].

Part_I_LipidCompositions/Supplemental_Table_A.3_Stability.xlsx

Table A.3: Stability of factors f describing PC compositions
The stability of factors f of PCs measured on the fatty acid level (Lipidyzer$^{\mathrm{TM}}$) and PCs measured on the lipid species level (Absolute*IDQ*$^{\mathrm{TM}}$) was demonstrated within the distributions of subjects (determined by ANOVA or Kruskal-Wallis test) and between the time points of baseline and intervention (determined by Wilcoxon signed-rank test) of various challenges. Distributions that were not significantly different ($\alpha = 0.01$) were labeled as stable. Due to a small number of instances, the null hypothesis could not be rejected during tests in connection to challenges. Supplemental Figure A.1 and Supplemental Table A.2 indicate that variation of factors f between subjects is larger than during challenges. This table is part of my prepared manuscript as well; Quell et al. [149].

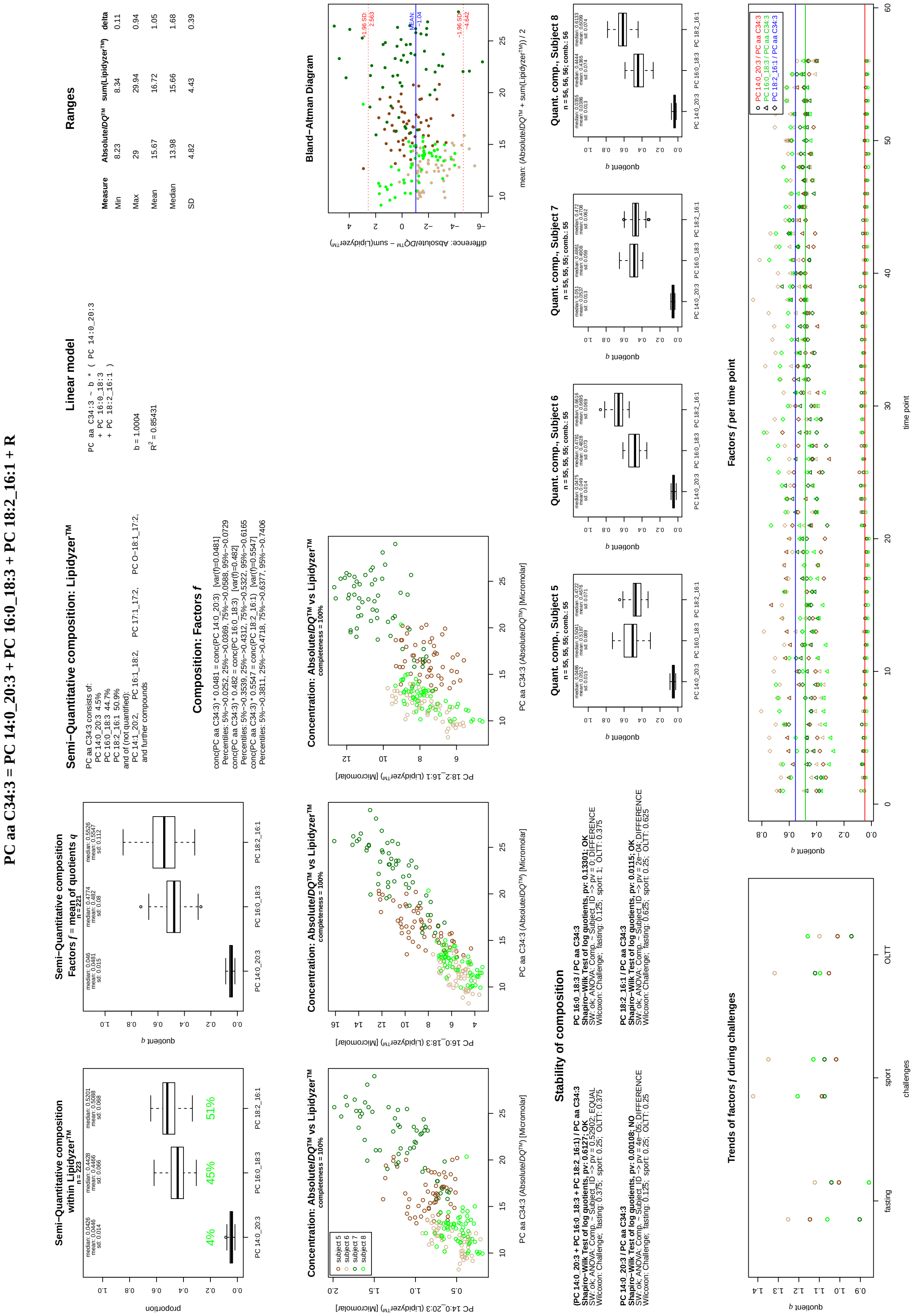
Part_I_LipidCompositions/Supplemental_Table_A.4_Replication.xlsx

Table A.4: Replication of factors f and imputation in QMDiab
Compositions of PCs measured the fatty acid level are replicated, if factors f (distributions of quotients q) of the major compound are not significantly different between subjects of the HuMet study and a comparable subset of QMDiab (male controls aged $\leq$40y). In addition, stability of factors f is demonstrated between heterogeneous subsets of QMDiab (male controls, female controls and cases). Furthermore, imputation is replicated if distributions of measured PC concentrations (fatty acid level) of the subset of QMDiab are not significantly different to concentrations of PC sums of the subset of QMDiab multiplied by the respective factor f determined in HuMet.

Part_I_LipidCompositions/Supplemental_Table_A.5_LinearModel.xlsx

Table A.5: Explained variation of PCs measured on the lipid species level
The variance of in median 67.2% (R^2, mean: 0.586, SD: 0.277) of PCs measured on the lipid species level (Absolute*IDQ*$^{\mathrm{TM}}$) can be explained by the sum of related PCs measured on the fatty acid level (Lipidyzer$^{\mathrm{TM}}$). This table is part of my prepared manuscript as well; Quell et al. [149].

Part_I_LipidCompositions/Supplemental_Figure_A.1_CompDetails.pdf
Exemplary excerpt:



PC aa C34:3 = PC 14:0_20:3 + PC 16:0_18:3 + PC 18:2_16:1 + R

Ranges

Measure	Absolute*IDQ*™	sum(Lipidyzer™)	delta
Min	8.23	8.34	0.11
Max	29	29.94	0.94
Mean	15.67	16.72	1.05
Median	13.98	15.66	1.68
SD	4.82	4.43	0.39

Linear model

```
PC aa C34:3 ~ b * ( PC 14:0_20:3
  + PC 16:0_18:3
  + PC 18:2_16:1 )
```

b = 1.0004

$R^2 = 0.85431$

Semi-Quantitative composition: Lipidyzer™

PC aa C34:3 consists of:
PC 14:0_20:3 4.5%
PC 16:0_18:3 44.7%
PC 18:2_16:1 50.9%
and of (not quantified):
PC 14:1_20:2, PC 16:1_18:2, PC 17:1_17:2, PC O−18:1_17:2,
and further compounds

Composition: Factors *f*

conc(PC aa C34:3) * 0.0481 = conc(PC 14:0_20:3) [var(f)=0.0481]
Percentiles: 5%−>0.0252, 25%−>0.0369, 75%−>0.0588, 95%−>0.0729
conc(PC aa C34:3) * 0.482 = conc(PC 16:0_18:3) [var(f)=0.482]
Percentiles: 5%−>0.3539, 25%−>0.4312, 75%−>0.5322, 95%−>0.6165
conc(PC aa C34:3) * 0.5547 = conc(PC 18:2_16:1) [var(f)=0.5547]
Percentiles: 5%−>0.3811, 25%−>0.4718, 75%−>0.6377, 95%−>0.7406

Stability of composition

(PC 14:0_20:3 + PC 16:0_18:3 + PC 18:2_16:1) / PC aa C34:3
Shapiro−Wilk Test of log quotients, pv: 0.6127; OK
SW: ok; ANOVA: Comp. ~ Subject_ID −> pv = 0.52902; EQUAL
Wilcoxon: Challenge; fasting: 0.375; sport: 0.25; OLTT: 0.375

PC 16:0_18:3 / PC aa C34:3
Shapiro−Wilk Test of log quotients, pv: 0.13301; OK
SW: ok; ANOVA: Comp. ~ Subject_ID −> pv = 0; DIFFERENCE
Wilcoxon: Challenge; fasting: 0.125; sport: 1; OLTT: 0.375

PC 14:0_20:3 / PC aa C34:3
Shapiro−Wilk Test of log quotients, pv: 0.00108; NO
SW: ok; ANOVA: Comp. ~ Subject_ID −> pv = 4e−05; DIFFERENCE
Wilcoxon: Challenge; fasting: 0.125; sport: 0.25; OLTT: 0.25

PC 18:2_16:1 / PC aa C34:3
Shapiro−Wilk Test of log quotients, pv: 0.0115; OK
SW: ok; ANOVA: Comp. ~ Subject_ID −> pv = 2e−04; DIFFERENCE
Wilcoxon: Challenge; fasting: 0.625; sport: 0.25; OLTT: 0.625

Semi−Quantitative composition within Lipidyzer™ n = 223
proportion
4% 45% 51%
PC 14:0_20:3 PC 16:0_18:3 PC 18:2_16:1
Semi−Quantitative composition Factors f = mean of quotients q n = 221
quotient q
Concentration: Absolute/IDQ™ vs Lipidyzer™ completeness = 100%
PC 14:0_20:3 (Lipidyzer™) [Micromolar]
PC 16:0_18:3 (Lipidyzer™) [Micromolar]
PC 18:2_16:1 (Lipidyzer™) [Micromolar]
PC aa C34:3 (Absolute/IDQ™) [Micromolar]
subject 5 subject 6 subject 7 subject 8
Bland−Altman Diagram
+1.96 SD: 2.563
MEAN: −1.04
−1.96 SD: −4.642
difference: Absolute/IDQ™ − sum(Lipidyzer™)
mean: (Absolute/IDQ™ + sum(Lipidyzer™)) / 2
Quant. comp., Subject 5 n = 55, 55, 55; comb.: 55
Quant. comp., Subject 6 n = 55, 55, 55; comb.: 55
Quant. comp., Subject 7 n = 55, 55, 55; comb.: 55
Quant. comp., Subject 8 n = 56, 56, 56; comb.: 56
Trends of factors f during challenges
fasting sport OLTT
challenges
Factors f per time point
PC 14:0_20:3 / PC aa C34:3
PC 16:0_18:3 / PC aa C34:3
PC 18:2_16:1 / PC aa C34:3
time point

Figure A.1: Details of PC compositions
Collection of all plots and information of each measured PC composition (Chapter 3). A similar version of this figure is part of my prepared manuscript as well; Quell et al. [149].

Part_I_LipidCompositions/Supplemental_Figure_A.2_Replication_factor.pdf
Exemplary excerpt:

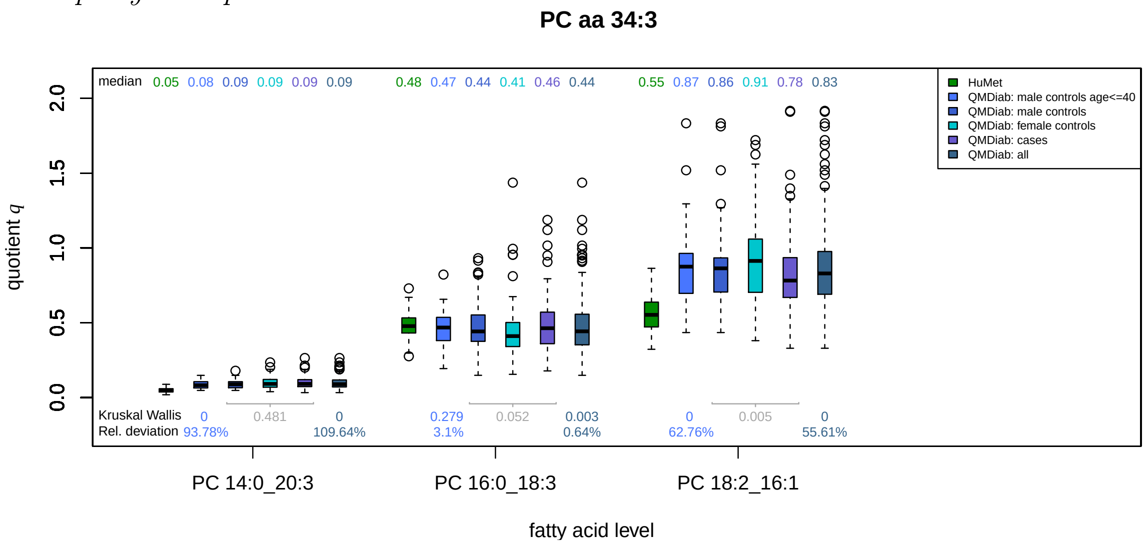


Figure A.2: Factors f in HuMet and QMDiab of PC compositions
Factors f (distributions of quotients q) are shown separately for subjects of HuMet and subsets of QMDiab. The p-values of Kruskal-Wallis tests indicate whether factors f of HuMet and of a comparable subset of QMDiab (male controls aged ≤ 40y) are significantly different. Further p-values refer to HuMet and complete QMDiab or to subsets of QMDiab (male controls, female controls, cases). This figure is part of my prepared manuscript as well; Quell et al. [149].

Part_I_LipidCompositions/Supplemental_Figure_A.3_Replication_conz.pdf
Exemplary excerpt:

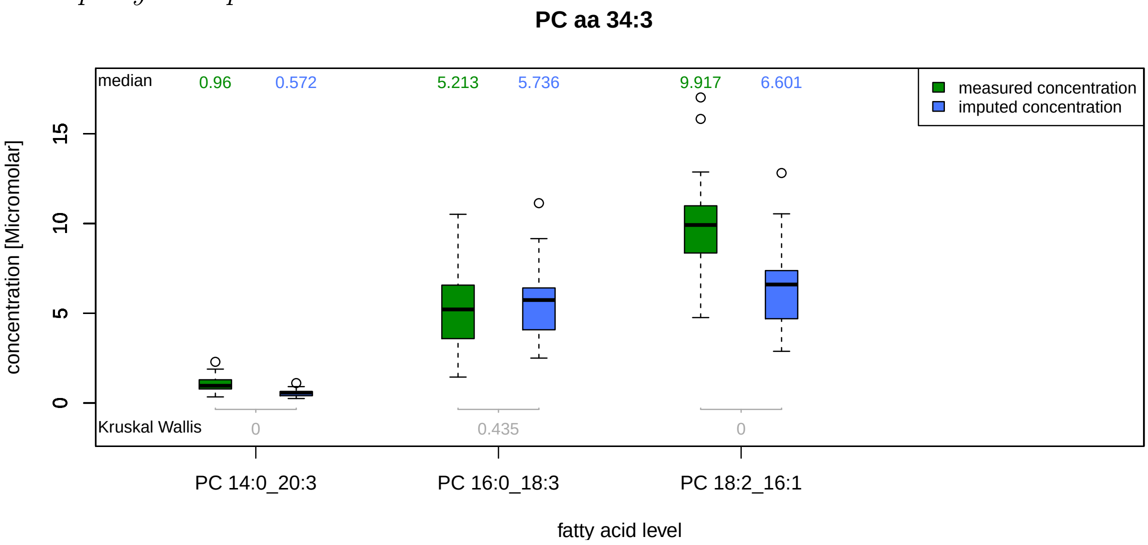


Figure A.3: Distributions of measured and imputed PC concentrations
PC concentrations (lipid species level) of a subset of QMDiab (male controls aged $\leq$ 40y) were multiplied by respective factors f determined in HuMet to impute concentrations of PCs of the fatty acid level. P-values of a Kruskal-Wallis test show whether distributions of measured and imputed concentrations are not significantly different ($\alpha = 0.01$).

APPENDIX B

Supplements: Annotation of unknown metabolites

Index of figures, tables, and files of Appendix B

Figures

Tables

Files

Automated data integration procedure

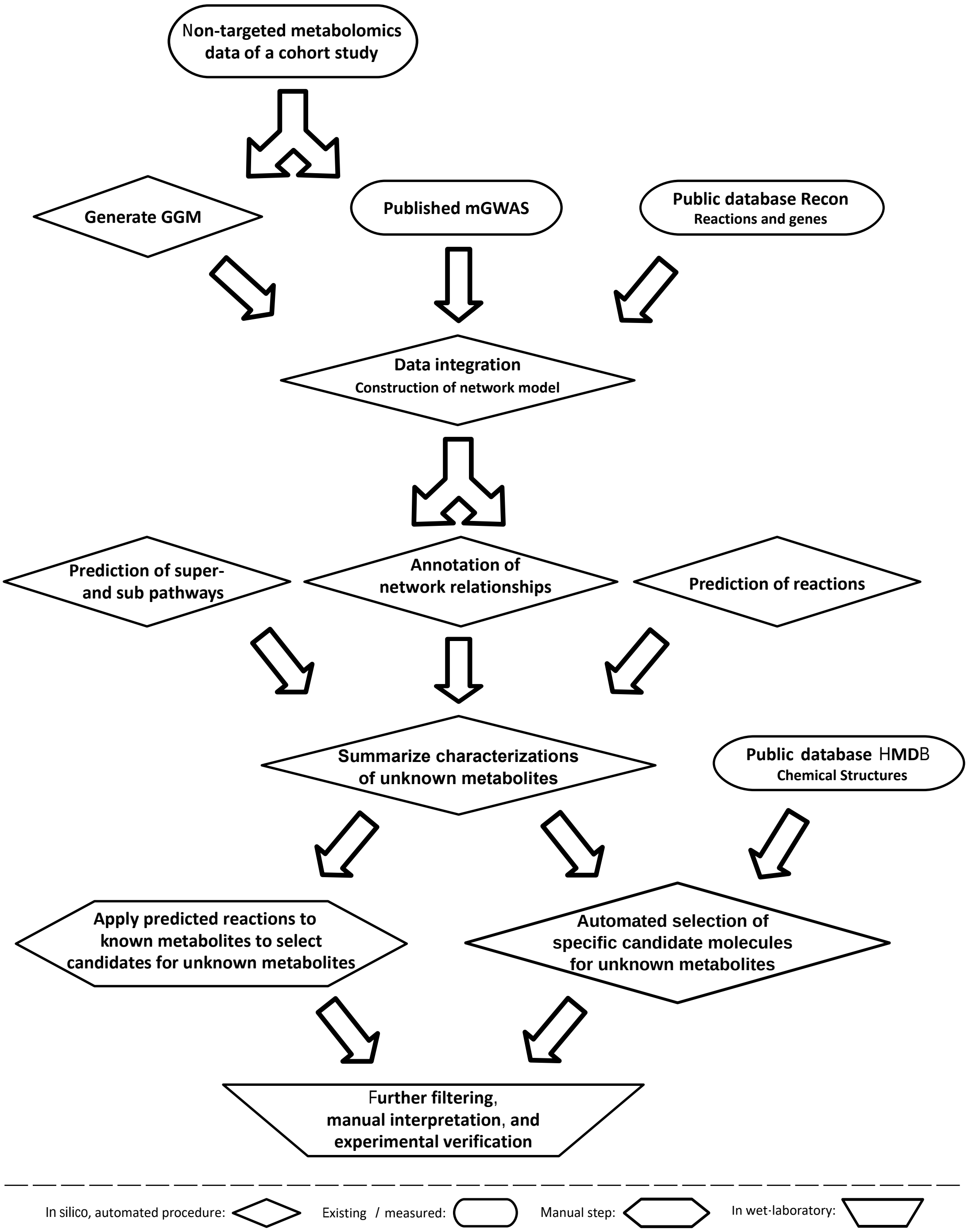

Non-targeted metabolomics data of a cohort study
Generate GGM
Published mGWAS
Public database Recon
Reactions and genes
Data integration
Construction of network model
Prediction of super- and sub pathways
Annotation of network relationships
Prediction of reactions
Summarize characterizations of unknown metabolites
Public database HMDB
Chemical Structures
Apply predicted reactions to known metabolites to select candidates for unknown metabolites
Automated selection of specific candidate molecules for unknown metabolites
Further filtering, manual interpretation, and experimental verification
In silico, automated procedure:
Existing / measured:
Manual step:
In wet-laboratory:

Figure B.1: Basic elements of the unknown characterization procedure
The schematic workflow shows the sequential steps of the introduced procedure to characterize unknown metabolites. The shape of each element indicates the respective part consisting of an automated *in silico* procedure, existing data, manual, or wet laboratory work. I previously published a similar version this figure in Quell et al. 2017; adapted by permission from the Journal of Chromatography B [150].

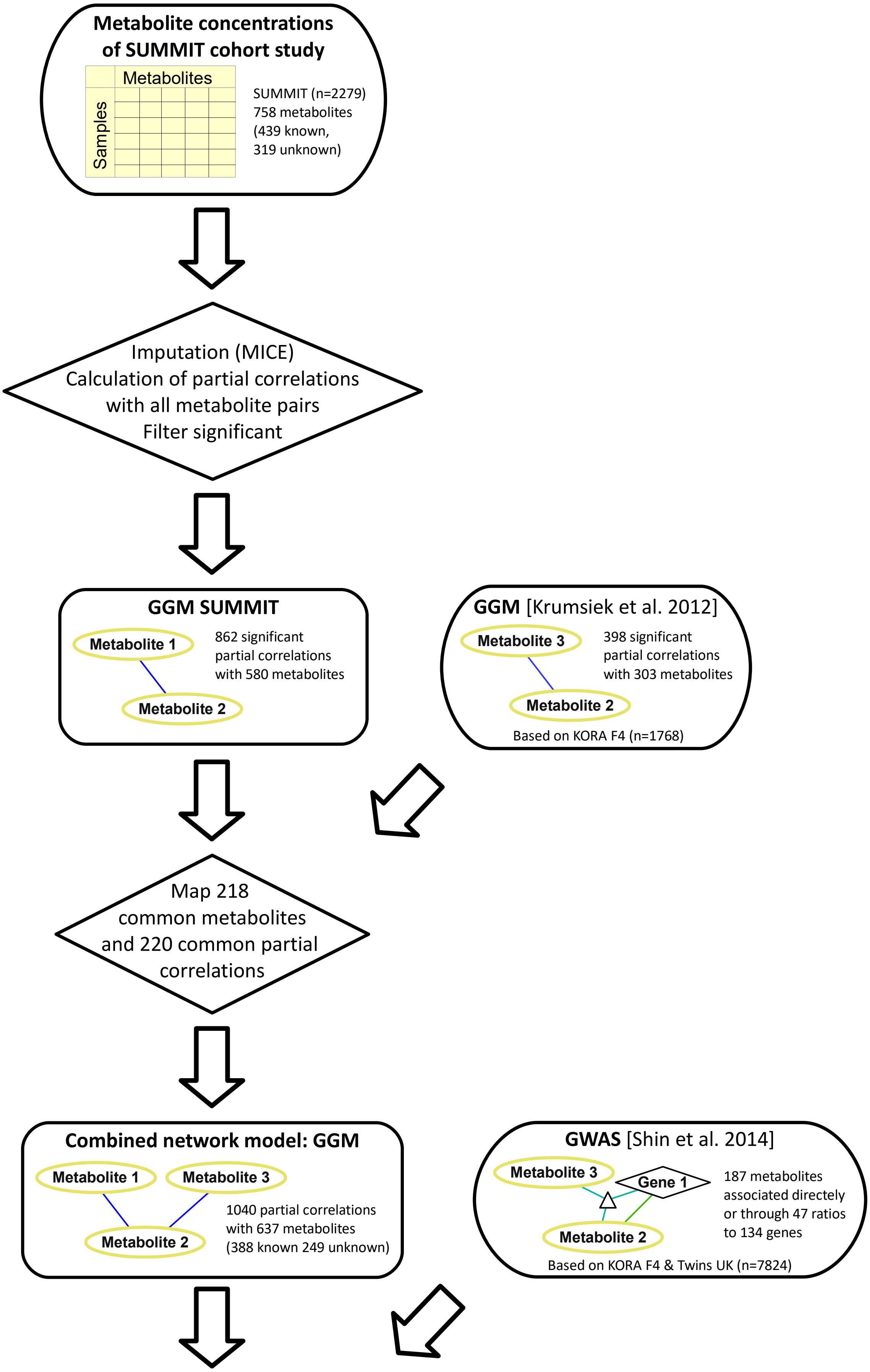

Metabolite concentrations of SUMMIT cohort study
Metabolites
Samples
SUMMIT (n=2279)
758 metabolites
(439 known,
319 unknown)
Imputation (MICE)
Calculation of partial correlations
with all metabolite pairs
Filter significant
GGM SUMMIT
Metabolite 1
Metabolite 2
862 significant
partial correlations
with 580 metabolites
GGM [Krumsiek et al. 2012]
Metabolite 3
Metabolite 2
398 significant
partial correlations
with 303 metabolites
Based on KORA F4 (n=1768)
Map 218
common metabolites
and 220 common partial
correlations
Combined network model: GGM
Metabolite 1
Metabolite 3
Metabolite 2
1040 partial correlations
with 637 metabolites
(388 known 249 unknown)
GWAS [Shin et al. 2014]
Metabolite 3
Gene 1
Metabolite 2
187 metabolites
associated directely
or through 47 ratios
to 134 genes
Based on KORA F4 & Twins UK (n=7824)

Figure continued (p. 2/2)

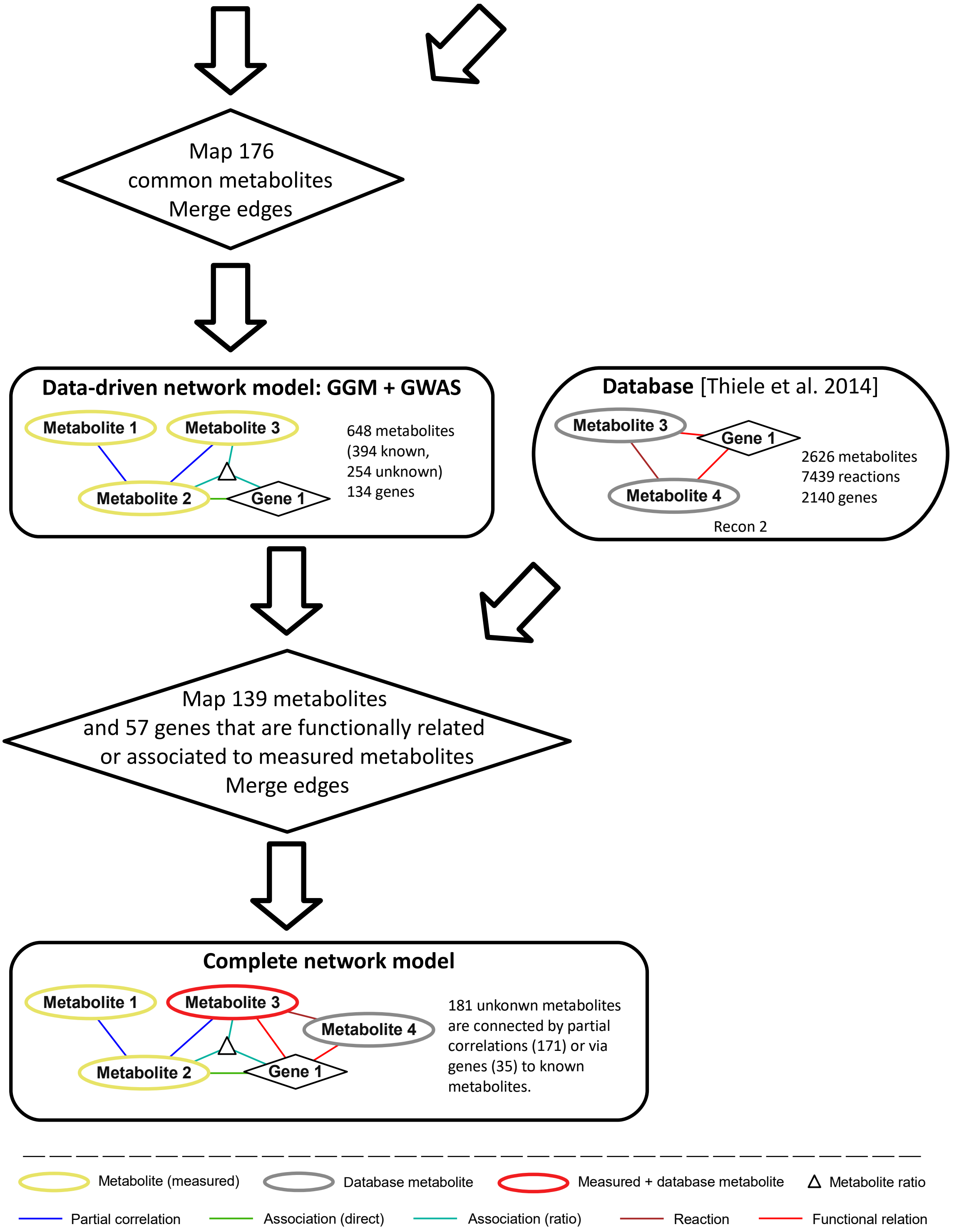


Figure B.2: Network model: stepwise data integration procedure (SUMMIT)
The network model is composed by three several types of data that are integrated during sequential steps. The workflow describes each step together including its input, output, and intermediate states of the network model. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Automated annotation of unknown metabolites

Confidence	Total	Correct	Similar	False	Ff	None
very high	10 797	8 354 (77.4%)	953 (8.8%)	1 317 (12.1%)	183 (1.7%)	0 (0.0%)
high	1 278	682 (53.4%)	0 (0.0%)	533 (41.7%)	63 (4.9%)	0 (0.0%)
medium	9 914	5 932 (59.8%)	1 302 (13.1%)	1 781 (18.0%)	899 (9.1%)	0 (0.0%)
low	5 308	1 957 (36.9%)	93 (1.8%)	1 941 (36.6%)	1 218 (22.9%)	99 (1.9%)
very low	2 985	827 (27.7%)	305 (10.2%)	691 (23.1%)	1 162 (38.9%)	0 (0.0%)

Table B.1: Cross-validation of the pathway prediction module
The pathway prediction module went through 100x 10-fold cross-validation receiving the total count of predicted super pathways per confidence level. Following columns refer to the number of accompanying predicted sub pathways. For the sub pathway prediction, we distinguished between correct, similar (e.g. long-chain fatty acids versus medium-chain fatty acids), false, false because of wrong super pathway (Ff), or no prediction. I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Name	Predicted super pathway	Confidence	Predicted sub pathway	Fraction
X-11805	peptide	very high	dipeptide	0.8
X-13429	lipid	very high	sterol, steroid	0.4
X-12063	lipid	very high	sterol, steroid	0.4
X-11440	lipid	very high	sterol, steroid	0.6
X-12456	lipid	very high	sterol, steroid	0.5
X-11792	peptide	very high	dipeptide	0.8
X-12850	lipid	very high	sterol, steroid	0.6
X-14086	peptide	very high	dipeptide	1
X-11441	cofactors and vitamins	very high	hemoglobin and porphyrin metabolism	1
X-11442	cofactors and vitamins	very high	hemoglobin and porphyrin metabolism	1
X-11530	cofactors and vitamins	very high	hemoglobin and porphyrin metabolism	1
X-17174	peptide	very high	dipeptide	0.7
X-18601	lipid	very high	sterol, steroid	0.8
X-08988	amino acid	very high	glycine, serine and threonine metabolism	0.6
X-11381	lipid	very high	carnitine metabolism	0.6
X-11438	lipid	very high	fatty acid, dicarboxylate, or long chain fatty acid	0.4
X-11491	lipid	very high	fatty acid, dicarboxylate, or lysolipid	0.3
X-11538	lipid	very high	fatty acid, dicarboxylate	0.5
X-11469	lipid	very high	sterol, steroid	0.5
X-12644	lipid	very high	lysolipid	0.8
X-11529	lipid	very high	fatty acid, dicarboxylate, or lysolipid	0.4
X-14626	lipid	very high	fatty acid, dicarboxylate, or lysolipid	0.4
X-02249	lipid	very high	essential fatty acid, or fatty acid, dicarboxylate, or long chain fatty acid	0.3
X-11421	lipid	very high	carnitine metabolism	0.7
X-11470	lipid	very high	sterol/steroid	0.7
X-17269	lipid	very high	medium chain fatty acid	1
X-14192	peptide	very high	dipeptide	1
X-14272	peptide	very high	dipeptide, or gamma-glutamyl amino acid	0.5
X-16123	peptide	very high	dipeptide	1
X-16128	peptide	very high	dipeptide	1
X-16132	peptide	very high	dipeptide	1
X-16134	peptide	very high	fibrinogen cleavage peptide	1
X-12556	amino acid	very high	glycine, serine and threonine metabolism	0.8
X-11422	nucleotide	high	purine metabolism, (hypo)xanthine/inosine containing	0.7
X-13435	lipid	high	carnitine metabolism	1
X-11261	lipid	high	carnitine metabolism	0.4
X-12798	lipid	high	carnitine metabolism	0.7
X-02269	lipid	medium	fatty acid, dicarboxylate, or long chain fatty acid	0.5
X-08402	lipid	medium	sphingolipid, or sterol/steroid	0.5
X-10510	lipid	medium	sphingolipid, or sterol/steroid	0.5
X-11244	lipid	medium	sterol, steroid	1
X-11443	lipid	medium	sterol, steroid	1
X-11820	lipid	medium	carnitine metabolism, or sterol/steroid	0.5
X-11905	lipid	medium	fatty acid, dicarboxylate	1
X-12450	lipid	medium	essential fatty acid, or fatty acid, monohydroxy	0.5

Table continued (p. 2/5)

Name	Predicted super pathway	Confidence	Predicted sub pathway	Fraction
X-12465	lipid	medium	carnitine metabolism, or ketone bodies	0.5
X-12627	lipid	medium	essential fatty acid, or long chain fatty acid	0.5
X-13891	lipid	medium	fatty acid, dicarboxylate	1
X-14632	lipid	medium	bile acid metabolism, or sterol/steroid	0.5
X-14658	lipid	medium	bile acid metabolism	1
X-16654	lipid	medium	bile acid metabolism	1
X-17443	lipid	medium	fatty acid, monohydroxy	1
X-11522	cofactors and vitamins	medium	hemoglobin and porphyrin metabolism	1
X-04495	amino acid	medium	butanoate metabolism, or creatine metabolism, or cysteine, methionine, sam, taurine metabolism	0.3
X-09706	amino acid	medium	urea cycle; arginine-, proline-, metabolism, or valine, leucine and isoleucine metabolism	0.5
X-11478	amino acid	medium	phenylalanine & tyrosine metabolism, or tryptophan metabolism	0.5
X-11837	amino acid	medium	phenylalanine & tyrosine metabolism	1
X-12216	amino acid	medium	phenylalanine & tyrosine metabolism	1
X-14352	amino acid	medium	urea cycle; arginine-, proline-, metabolism, or valine, leucine and isoleucine metabolism	0.5
X-11838	xenobiotics	medium	drug	1
X-12039	xenobiotics	medium	food component/plant, or xanthine metabolism	0.5
X-14374	xenobiotics	medium	benzoate metabolism, or xanthine metabolism	0.5
X-11429	nucleotide	medium	purine metabolism, (hypo)xanthine/inosine containing, or pyrimidine metabolism, uracil containing	0.5
X-16674	lipid	medium	fatty acid, monohydroxy	1
X-03094	lipid	medium	sterol/steroid	1
X-10419	lipid	medium	sterol/steroid	1
X-10500	lipid	medium	sterol/steroid	1
X-11247	lipid	medium	long chain fatty acid	1
X-11317	lipid	medium	lysolipid	1
X-11327	lipid	medium	carnitine metabolism	1
X-11450	lipid	medium	sterol, steroid	1
X-11508	lipid	medium	fatty acid, monohydroxy	1
X-11521	lipid	medium	essential fatty acid	1
X-11533	lipid	medium	medium chain fatty acid	1
X-11537	lipid	medium	glycerolipid metabolism	1
X-11540	lipid	medium	glycerolipid metabolism	1
X-11550	lipid	medium	medium chain fatty acid	1
X-11552	lipid	medium	fatty acid, amide	1
X-11859	lipid	medium	medium chain fatty acid	1
X-12051	lipid	medium	lysolipid	1
X-13069	lipid	medium	long chain fatty acid	1
X-14662	lipid	medium	bile acid metabolism	1
X-14939	lipid	medium	medium chain fatty acid	1
X-15222	lipid	medium	medium chain fatty acid	1
X-15492	lipid	medium	sterol/steroid	1
X-16578	lipid	medium	medium chain fatty acid	1
X-16943	lipid	medium	medium chain fatty acid	1

Table continued (p. 3/5)

Name	Predicted super pathway	Confidence	Predicted sub pathway	Fraction
X-16947	lipid	medium	inositol metabolism	1
X-17254	lipid	medium	lysolipid	1
X-17299	lipid	medium	carnitine metabolism	1
X-17438	lipid	medium	fatty acid, dicarboxylate	1
X-06307	peptide	medium	dipeptide	1
X-12038	peptide	medium	polypeptide	1
X-16130	peptide	medium	dipeptide	1
X-16135	peptide	medium	fibrinogen cleavage peptide	1
X-16137	peptide	medium	polypeptide	1
X-17189	peptide	medium	dipeptide	1
X-17441	peptide	medium	fibrinogen cleavage peptide	1
X-14095	amino acid	medium	—	
X-11333	amino acid	medium	amino fatty acid, or lysine metabolism, or urea cycle; arginine-, proline-, metabolism	0.3
X-11809	cofactors and vitamins	low	hemoglobin and porphyrin metabolism	1
X-12206	cofactors and vitamins	low	ascorbate and aldarate metabolism	1
X-14056	cofactors and vitamins	low	hemoglobin and porphyrin metabolism	1
X-16124	cofactors and vitamins	low	hemoglobin and porphyrin metabolism	1
X-16946	cofactors and vitamins	low	hemoglobin and porphyrin metabolism	1
X-17162	cofactors and vitamins	low	hemoglobin and porphyrin metabolism	1
X-17612	cofactors and vitamins	low	hemoglobin and porphyrin metabolism	1
X-16480	lipid	low	essential fatty acid, or fatty acid, dicarboxylate	0.5
X-12093	amino acid	low	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	0.5
X-12511	amino acid	low	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	0.5
X-13477	amino acid	low	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	0.5
X-03088	amino acid	low	urea cycle; arginine-, proline-, metabolism	1
X-05491	amino acid	low	butanoate metabolism	1
X-06126	amino acid	low	phenylalanine & tyrosine metabolism	1
X-06246	amino acid	low	alanine and aspartate metabolism	1
X-06267	amino acid	low	urea cycle; arginine-, proline-, metabolism	1
X-11334	amino acid	low	lysine metabolism	1
X-11818	amino acid	low	amino fatty acid	1
X-12405	amino acid	low	tryptophan metabolism	1
X-12734	amino acid	low	phenylalanine & tyrosine metabolism	1
X-12749	amino acid	low	phenylalanine & tyrosine metabolism	1
X-12786	amino acid	low	alanine and aspartate metabolism	1
X-12802	amino acid	low	valine, leucine and isoleucine metabolism	1
X-13619	amino acid	low	urea cycle; arginine-, proline-, metabolism	1
X-13835	amino acid	low	histidine metabolism	1
X-14588	amino acid	low	lysine metabolism	1
X-15461	amino acid	low	phenylalanine & tyrosine metabolism	1
X-15667	amino acid	low	tryptophan metabolism	1
X-16071	amino acid	low	tryptophan metabolism	1
X-17138	amino acid	low	valine, leucine and isoleucine metabolism	1

Table continued (p. 4/5)

Name	Predicted super pathway	Confidence	Predicted sub pathway	Fraction
X-17685	amino acid	low	phenylalanine & tyrosine metabolism	1
X-12543	amino acid	low	phenylalanine & tyrosine metabolism	1
X-14625	amino acid	low	glutamate metabolism, or glutathione metabolism	0.5
X-13848	amino acid	low	—	
X-10810	nucleotide	low	purine metabolism, (hypo)xanthine/inosine containing	1
X-12094	nucleotide	low	nad metabolism	1
X-12844	nucleotide	low	nad metabolism	1
X-11452	xenobiotics	low	food component/plant	1
X-12040	xenobiotics	low	food component, plant	1
X-12217	xenobiotics	low	benzoate metabolism	1
X-12230	xenobiotics	low	benzoate metabolism	1
X-12231	xenobiotics	low	food component/plant	1
X-12329	xenobiotics	low	food component/plant	1
X-12407	xenobiotics	low	food component/plant	1
X-12730	xenobiotics	low	food component/plant	1
X-12816	xenobiotics	low	food component/plant	1
X-12830	xenobiotics	low	food component/plant	1
X-12847	xenobiotics	low	food component/plant	1
X-13728	xenobiotics	low	xanthine metabolism	1
X-15497	xenobiotics	low	drug	1
X-15728	xenobiotics	low	benzoate metabolism	1
X-16564	xenobiotics	low	benzoate metabolism	1
X-16940	xenobiotics	low	benzoate metabolism	1
X-17150	xenobiotics	low	food component/plant	1
X-17185	xenobiotics	low	benzoate metabolism	1
X-17314	xenobiotics	low	benzoate metabolism	1
X-12544	lipid	very low	sterol/steroid	1
X-02973	cofactors and vitamins	very low	ascorbate and aldarate metabolism	1
X-04357	carbohydrate	very low	fructose, mannose, galactose, starch, and sucrose metabolism	1
X-12007	carbohydrate	very low	fructose, mannose, galactose, starch, and sucrose metabolism	1
X-12056	carbohydrate	very low	fructose, mannose, galactose, starch, and sucrose metabolism	1
X-12116	cofactors and vitamins	very low	ascorbate and aldarate metabolism	1
X-12696	carbohydrate	very low	glycolysis, gluconeogenesis, pyruvate metabolism	1
X-13727	carbohydrate	very low	fructose, mannose, galactose, starch, and sucrose metabolism	1
X-17502	carbohydrate	very low	glycolysis, gluconeogenesis, pyruvate metabolism	1
X-18221	carbohydrate	very low	glycolysis, gluconeogenesis, pyruvate metabolism	1
X-14473	peptide	very low	dipeptide	1
X-11799	lipid	very low	inositol metabolism	1
X-10506	amino acid	very low	alanine and aspartate metabolism	1
X-11315	amino acid	very low	glutamate metabolism	1
X-17145	amino acid	very low	tryptophan metabolism	1
X-15245	carbohydrate	very low	glycolysis, gluconeogenesis, pyruvate metabolism	1

Table continued (p. 5 / 5)

Name	Predicted super pathway	Confidence	Predicted sub pathway	Fraction
X-11561	peptide	very low	dipeptide	1
X-14302	peptide	very low	polypeptide	1
X-01911	cofactors and vitamins	very low	ascorbate and aldarate metabolism	1
X-11319	lipid	very low	long chain fatty acid	1
X-13866	lipid	very low	long chain fatty acid	1
X-11787	amino acid	very low	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	0.5
X-11255	amino acid	very low	valine, leucine and isoleucine metabolism	1
X-12704	amino acid	very low	phenylalanine & tyrosine metabolism	1

Table B.2: Predicted super and sub pathways of unknown metabolites
Super and sub pathways were automatically predicted for 180 and 178 unknown metabolites, respectively. Predicted sub pathways come along with their fraction among neighboring metabolites within the network model (in case of more than one predicted sub pathway, factions of the most frequent one is specified). Entries are ordered by their confidence scores decreasingly. I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Metabolite	Unknown	P. super pw	P. sub pw	$\triangle$mass	Predicted reaction
11-HpODE	X-11421			0.98	Oxidative deamination
2-methylbutyroyl-carnitine	X-11255	amino acid	valine, leucine and isoleucine metabolism	1.0*	Oxidative deamination
pyroglutamine	X-11315	amino acid	glutamate metabolism	1.0*	Oxidative deamination
3-methylhistidine	X-13835	amino acid	histidine metabolism	1.0*	Oxidative deamination
X-14632	X-14658			1.0	Oxidative deamination
Guanine	X-11422			1.1	Oxidative deamination
X-11244	X-11443			1.1	Oxidative deamination
X-11443	X-11450			-1.1	Oxidative deamination
N-acetylornithine	X-13477	amino acid	urea cycle; arginine-, proline-, metabolism	1.1	Oxidative deamination
X-11378	X-16935			1.8	De-hydrogenation/ Reduction
X-12217	X-16940			1.8	De-hydrogenation/ Reduction
X-11444	X-12844			-1.9	De-hydrogenation/ Reduction
X-12230	X-17185			-1.9	De-hydrogenation/ Reduction
X-11444	X-17706			-1.9	De-hydrogenation/ Reduction
pelargonate (9:0)	X-11859	lipid	medium chain fatty acid	-1.9	De-hydrogenation/ Reduction
decanoylcarnitine	X-13435	lipid	carnitine metabolism	-1.9	De-hydrogenation/ Reduction
estrone	X-02249			-1.9	De-hydrogenation/ Reduction
pseudouridine	X-11429	nucleotide	pyrimidine metabolism, uracil containing	2.0*	De-hydrogenation/ Reduction
3-carboxy-4-methyl-5-propyl 2-furanpropanoate (CMPF)	X-11469	lipid	fatty acid, dicarboxylate	-2.0*	De-hydrogenation/ Reduction
X-11299	X-11483			-2.0*	De-hydrogenation/ Reduction
octadecanedioate	X-11538	lipid	fatty acid, dicarboxylate	-2.0*	De-hydrogenation/ Reduction
hexadecanedioate	X-11905	lipid	fatty acid, dicarboxylate	-2.0*	De-hydrogenation/ Reduction
thymol sulfate	X-12847	xenobiotics	food component/plant	-2.0*	De-hydrogenation/ Reduction
X-11538	X-16480			-2.0	De-hydrogenation/ Reduction
L-urobilin	X-17162	cofactors and vitamins	hemoglobin and porphyrin metabolism	-2.0*	De-hydrogenation/ Reduction
X-12844	X-17357			2.0	De-hydrogenation/ Reduction
X-12846	X-17703			-2.0	De-hydrogenation/ Reduction
X-15492	X-17706			-2.0	De-hydrogenation/ Reduction
omega hydroxy tetradecanoate (n C14:0)	X-11438			-2.0	De-hydrogenation/ Reduction
pelargonate (9:0)	X-17269	lipid	medium chain fatty acid	-2.0	De-hydrogenation/ Reduction
dehydroisoandrosterone sulfate (DHEA-S)	X-18601	lipid	sterol, steroid	2.0	De-hydrogenation/ Reduction
4-hydroxyphenyl-pyruvate	X-12543	amino acid	phenylalanine & tyrosine metabolism	2.1	De-hydrogenation/ Reduction
dodecanedioate	X-13891	lipid	fatty acid, dicarboxylate	-2.1*	De-hydrogenation/ Reduction
X-17359	X-17706			-2.1	De-hydrogenation/ Reduction
5,8-tetradecadienoate	X-13069	lipid	long chain fatty acid	2.3*	De-hydrogenation/ Reduction
myristate (14:0)	X-11438	lipid	long chain fatty acid	14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
X-11317	X-11497			14.0	Methylation/ De-methylation, or Alkyl-chain-elongation

Table continued (p. 2/5)

Metabolite	Unknown	P. super pw	P. sub pw	△mass	Predicted reaction
13-cis-retinoate	X-11530			14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
all-trans-retinoate	X-11530			14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
X-14374	X-14473			14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
X-12212	X-15636			14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
X-11441	X-16946			-14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
X-12734	X-17685			14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
palmitate (16:0)	X-11438	lipid	long chain fatty acid	-14.0	Methylation/ De-methylation, or Alkyl-chain-elongation
estrone	X-02269			-14.1	Methylation/ De-methylation, or Alkyl-chain-elongation
8-hydroxyoctanoate	X-11508	lipid	fatty acid, monohydroxy	14.1*	Methylation/ De-methylation, or Alkyl-chain-elongation
2-aminooctanoic acid	X-11818	amino acid	amino fatty acid	-14.1*	Methylation/ De-methylation, or Alkyl-chain-elongation
X-12039	X-12329			14.1	Methylation/ De-methylation, or Alkyl-chain-elongation
X-11470	X-12844			14.1	Methylation/ De-methylation, or Alkyl-chain-elongation
X-02249	X-13866			-14.1	Methylation/ De-methylation, or Alkyl-chain-elongation
catechol sulfate	X-12217	xenobiotics	benzoate metabolism	14.2*	Methylation/ De-methylation, or Alkyl-chain-elongation
X-12734	X-16940			-14.2	Methylation/ De-methylation, or Alkyl-chain-elongation
urate	X-11422	nucleotide	purine metabolism, urate metabolism	-15.9	Oxidation, or Hydroxylation, or Epoxidation
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	X-02269	lipid	fatty acid, dicarboxylate	16.0*	Oxidation, or Hydroxylation, or Epoxidation
p-cresol sulfate	X-06126	amino acid	phenylalanine & tyrosine metabolism	16.0*	Oxidation, or Hydroxylation, or Epoxidation
tetradecanedioate	X-11438	lipid	fatty acid, dicarboxylate	-16.0*	Oxidation, or Hydroxylation, or Epoxidation
X-11444	X-11470			-16.0	Oxidation, or Hydroxylation, or Epoxidation
4-ethylphenylsulfate	X-12230	xenobiotics	benzoate metabolism	16.0*	Oxidation, or Hydroxylation, or Epoxidation
X-12329	X-12730			16.0	Oxidation, or Hydroxylation, or Epoxidation
X-12217	X-12734			16.0	Oxidation, or Hydroxylation, or Epoxidation
2-aminooctanoic acid	X-13477	amino acid	amino fatty acid	16.0*	Oxidation, or Hydroxylation, or Epoxidation
gamma glutamylglutamate	X-14272	peptide	gamma-glutamyl amino acid	-16.0*	Oxidation, or Hydroxylation, or Epoxidation
catechol sulfate	X-16940	xenobiotics	benzoate metabolism	16.0*	Oxidation, or Hydroxylation, or Epoxidation

Table continued (p. 3/5)

Metabolite	Unknown	P. super pw	P. sub pw	△mass	Predicted reaction
hypoxanthine	X-11422	nucleotide	purine metabolism, (hypo)xanthine/inosine containing	16.1	Oxidation, or Hydroxylation, or Epoxidation
estradiol	X-02269			-16.1	Oxidation, or Hydroxylation, or Epoxidation
3-phenylpropionate (hydrocinnamate)	X-11478	amino acid	phenylalanine & tyrosine metabolism	16.1*	Oxidation, or Hydroxylation, or Epoxidation
5alpha-androstan-3alpha, 17beta-diol disulfate	X-12544	lipid	sterol/steroid	-16.1*	Oxidation, or Hydroxylation, or Epoxidation
theobromine	X-14374	xenobiotics	xanthine metabolism	16.1*	Oxidation, or Hydroxylation, or Epoxidation
X-11470	X-15492			16.1	Oxidation, or Hydroxylation, or Epoxidation
4-vinylphenol sulfate	X-17185	xenobiotics	benzoate metabolism	16.1*	Oxidation, or Hydroxylation, or Epoxidation
linoleate (18:2n6)	X-12450	lipid	essential fatty acid	-27.8	De-ethylation, or Alkyl-chain-elongation
docosapentaenoate (n3 DPA; 22:5n3)	X-12627	lipid	essential fatty acid	28.0*	De-ethylation, or Alkyl-chain-elongation
X-12855	X-12860			28.0	De-ethylation, or Alkyl-chain-elongation
X-16674	X-17438			28.0	De-ethylation, or Alkyl-chain-elongation
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	X-02249	lipid	fatty acid, dicarboxylate	28.1*	De-ethylation, or Alkyl-chain-elongation
X-10346	X-11437			-28.1	De-ethylation, or Alkyl-chain-elongation
X-11538	X-11905			-28.1	De-ethylation, or Alkyl-chain-elongation
N-acetylornithine	X-12093	amino acid	urea cycle; arginine-, proline-, metabolism	28.1	De-ethylation, or Alkyl-chain-elongation
3-methylglutarylcarnitine (C6)	X-12802	amino acid	valine, leucine and isoleucine metabolism	28.1*	De-ethylation, or Alkyl-chain-elongation
X-11787	X-13477			28.1	De-ethylation, or Alkyl-chain-elongation
X-15728	X-16124			-28.1	De-ethylation, or Alkyl-chain-elongation
X-16940	X-17685			28.2	De-ethylation, or Alkyl-chain-elongation
2-hydroxyacetaminophen sulfate	X-11838	xenobiotics	drug	29.9*	Quinone, or CH3 to COOH, or Nitro reduction
4-ethylphenylsulfate	X-15728	xenobiotics	benzoate metabolism	30.0*	Quinone, or CH3 to COOH, or Nitro reduction
omega hydroxy hexadecanoate (n C16:0)	X-11438			-30.0	Quinone, or CH3 to COOH, or Nitro reduction
16alpha-Hydroxyestrone	X-02269			-30.1	Quinone, or CH3 to COOH, or Nitro reduction
X-02249	X-11469			-30.1	Quinone, or CH3 to COOH, or Nitro reduction
13-cis-retinoate	X-11441			31.9	Bis-oxidation
all-trans-retinoate	X-11441			31.9	Bis-oxidation

Table continued (p. 4/5)

Metabolite	Unknown	P. super pw	P. sub pw	△mass	Predicted reaction
13-cis-retinoate	X-11442			31.9	Bis-oxidation
all-trans-retinoate	X-11442			31.9	Bis-oxidation
propionylcarnitine	X-11381	lipid	fatty acid metabolism (also bcaa metabolism)	-31.9	Bis-oxidation
pregnenolone sulfate	X-12456	lipid	sterol/steroid	32.0	Bis-oxidation
estrone	X-11469			-32.1	Bis-oxidation
2-Hydroxyestradiol-17beta	X-02269			-32.1	Bis-oxidation
linolenate [α or γ; (18:3n3 or 6)]	X-16480	lipid	essential fatty acid	32.1	Bis-oxidation
lathosterol	X-12063	lipid	sterol/steroid	41.9	Acetylation
lathosterol	X-12456	lipid	sterol/steroid	41.9	Acetylation
5alpha-cholest-8-en-3beta-ol	X-12063			41.9	Acetylation
5alpha-cholest-8-en-3beta-ol	X-12456			41.9	Acetylation
androsterone	X-11441			41.9	Acetylation
androsterone	X-11442			41.9	Acetylation
carnosine	X-11561			42.0	Acetylation
X-12253	X-12258			42.0	Acetylation
2-aminooctanoic acid	X-12511	amino acid	amino fatty acid	42.0*	Acetylation
X-12802	X-12860			-42.0	Acetylation
Isocitrate	X-15245			42.0	Acetylation
estradiol	X-11530			42.0	Acetylation
laurate	X-11438			42.0	Acetylation
1-docosahexaenoylglycero-phosphocholine	X-12644	lipid	lysolipid	-42.1*	Acetylation
aflatoxin B1 exo-8,9-epoxide	X-18601			42.1	Acetylation
cholesta-5,7-dien-3beta-ol	X-12063			43.9	Decarboxylation
cholesta-5,7-dien-3beta-ol	X-12456			43.9	Decarboxylation
5alpha-cholesta-7,24-dien-3beta-ol	X-12063			43.9	Decarboxylation
5alpha-cholesta-7,24-dien-3beta-ol	X-12456			43.9	Decarboxylation
testosterone	X-11441			43.9	Decarboxylation
testosterone	X-11442			43.9	Decarboxylation
glucose	X-12007	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism	43.9	Decarboxylation
hexadecanedioate	X-11438	lipid	fatty acid, dicarboxylate	-44.0*	Decarboxylation
andro steroid monosulfate 2	X-12063	lipid	sterol/steroid	44.0*	Decarboxylation
oxalatosuccinate(3-)	X-15245			44.0	Decarboxylation
estrone	X-11530			44.0	Decarboxylation
acetylcarnitine	X-12465	lipid	carnitine metabolism	44.1	Decarboxylation

Table continued (p. 5/5)

Metabolite	Unknown	P. super pw	P. sub pw	$\triangle$mass	Predicted reaction
sebacate (decanedioate)	X-17438	lipid	fatty acid, dicarboxylate	44.1	Decarboxylation
X-13891	X-17443			44.1*	Decarboxylation
3-hydroxyhippurate	X-12704	xenobiotics	benzoate metabolism	79.9*	Sulfation, or Phosphatation
3-dehydrocarnitine	X-12798	lipid	carnitine metabolism	79.9*	Sulfation, or Phosphatation
4 androsten 3beta,17-beta-diol disulfate 2	X-18601	lipid	sterol, steroid	80.0	Sulfation, or Phosphatation
X-06126	X-11837			80.0	Sulfation, or Phosphatation
X-11308	X-11378			80.1	Sulfation, or Phosphatation
X-11378	X-17654			-80.1	Sulfation, or Phosphatation
X-14625	X-18221			95.8	Sulfation, or Phosphatation
caproate (6:0)	X-16578	lipid	medium chain fatty acid	95.9*	Sulfation, or Phosphatation
p-cresol sulfate	X-11837	amino acid	phenylalanine & tyrosine metabolism	96.0*	Sulfation, or Phosphatation
alpha-glutamyl-tyrosine	X-11805	peptide	dipeptide	107.0*	Taurine Conjugation
X-12830	X-17703			107.1	Taurine Conjugation
X-11261	X-11478			-119.0	Cys Conjugation
S-methylcysteine	X-13866	amino acid	cysteine, methionine, sam, taurine metabolism	119.1*	Cys Conjugation
dehydroisoandrosterone sulfate (DHEA-S)	X-11440	lipid	sterol, steroid	-120.9	Cys Conjugation
testosterone sulfate	X-11440			-120.9	Cys Conjugation
deoxycholate	X-11491	lipid	bile acid metabolism	176.0	Glucuronidation
1-arachidonoylglycerophosphoinositol	X-12063	lipid	lysolipid	-192.2*	Glucuronidation
1-arachidonoylglycerophosphoinositol	X-12456	lipid	lysolipid	-192.2*	Glucuronidation
X-09789	X-18774			192.2	Glucuronidation
phenylalanyl-phenylalanine	X-17189	peptide	dipeptide	306.8*	GSH Conjugation

Table B.3: Predicted reactions based on mass differences and network relations 100 reactions connecting known to unknown and 45 reactions among unknown metabolites were automatically selected based on their $\triangle$mass (± 0.3). Pairs of metabolites refer to neighborships in the network model via GGM edges, a common mGWAS gene, or a gene or third metabolite that is functionally related to a metabolite and associated to an unknown metabolite. The signs in the mass differences column indicate the direction of predicted reactions, starting with the (known) metabolite of the first column. An asterisk indicates that $\triangle$mass is based on the measured mass of known metabolites. The list is ordered by absolute $\triangle$mass values. I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

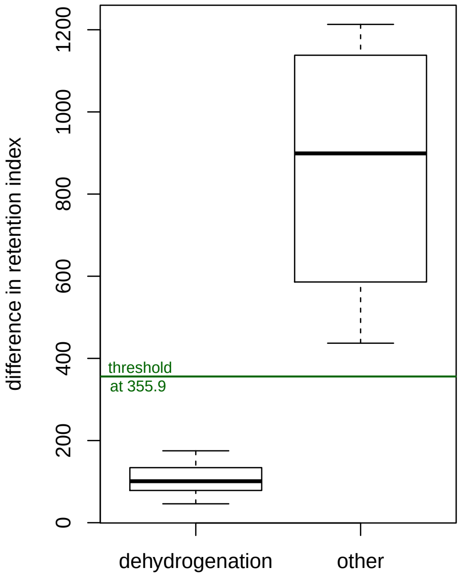


Figure B.3: Difference in retention index of dehydrogenation reactants
RIs of metabolite neighbors (according to their location in the network model) differ clearly between dehydrogenation reactants and random-pairs (Δmass = 2 ± 0.3). The threshold was determined by the mean of both distributions weighted by their SDs. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Metabolite	Unknown	Δmass	ΔRI	Predicted reaction
pelargonate (9:0)	X-11859	-1.9	294	De-hydrogenation/ Reduction
decanoylcarnitine	X-13435	-1.9	60	De-hydrogenation/ Reduction
pseudouridine	X-11429	2.0	47	De-hydrogenation/ Reduction
octadecanedioate	X-11538	-2.0	113	De-hydrogenation/ Reduction
hexadecanedioate	X-11905	-2.0	249	De-hydrogenation/ Reduction
thymol sulfate	X-12847	-2.0	155	De-hydrogenation/ Reduction
L-urobilin	X-17162	-2.0	50	De-hydrogenation/ Reduction
pelargonate (9:0)	X-17269	-2.0	317	De-hydrogenation/ Reduction
dehydroisoandrosterone sulfate (DHEA-S)	X-18601	2.0	229	De-hydrogenation/ Reduction
4-hydroxyphenylpyruvate	X-12543	2.1	265	De-hydrogenation/ Reduction
dodecanedioate	X-13891	-2.1	274	De-hydrogenation/ Reduction
5,8-tetradecadienoate	X-13069	2.3	94	De-hydrogenation/ Reduction
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	X-11469	-2.0	NA	measured on different platforms
estrone	X-02249	-2.0	NA	non-measured metabolite
omega hydroxy tetradecanoate (n-C14:0)	X-11438	-2.0	NA	non-measured metabolite
X-11444	X-12844	-1.9	185	De-hydrogenation/ Reduction
X-12230	X-17185	-1.9	291	De-hydrogenation/ Reduction
X-11538	X-16480	-2.0	235	De-hydrogenation/ Reduction
X-12844	X-17357	2.0	73	De-hydrogenation/ Reduction
X-12846	X-17703	-2.0	70	De-hydrogenation/ Reduction
X-15492	X-17706	-2.0	7	De-hydrogenation/ Reduction
X-17359	X-17706	-2.1	84	De-hydrogenation/ Reduction
X-11378	X-16935	1.8	860	no
X-12217	X-16940	1.8	649	no
X-11444	X-17706	-1.9	708	no
X-11299	X-11483	-2.0	450	no

Table B.4: Predicted dehydrogenation/ reduction reactions
Dehydrogenation reactions were assigned to neighboring metabolites (of the network model) with mass difference 2 ± 0.3 and post-filtered according to $\Delta\text{RI} \leq 355.9$. The signs in the Δmass-column indicate the direction of the reaction, starting from the first metabolite. I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Comparative annotation of several metabolite sets

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-01911	amino acid	cofactors and vitamins	—
X-02249	lipid	lipid	amino acid
X-02269	lipid	lipid	-/-
X-02973	cofactors and vitamins	cofactors and vitamins	-/-
X-03088	amino acid	amino acid	-/-
X-03094	lipid	lipid	-/-
X-04357	carbohydrate	carbohydrate	-/-
X-04495	amino acid	amino acid	-/-
X-04498	—	—	-/-
X-05426	—	—	-/-
X-05491	amino acid	amino acid	-/-
X-05907	—	—	-/-
X-06126	xenobiotics	amino acid	-/-
X-06226	—	—	-/-
X-06227	energy	—	-/-
X-06246	amino acid	amino acid	-/-
X-06267	amino acid	amino acid	-/-
X-06307	peptide	peptide	-/-
X-08402	lipid	lipid	-/-
X-08988	amino acid	amino acid	-/-
X-09026	—	—	-/-
X-09706	amino acid	amino acid	-/-
X-09789	xenobiotics	—	xenobiotics
X-10346	—	—	-/-
X-10395	—	—	-/-
X-10419	lipid	lipid	-/-
X-10457	—	-/-	—
X-10500	lipid	lipid	-/-
X-10506	amino acid	amino acid	-/-
X-10510	lipid	lipid	-/-
X-10810	nucleotide	nucleotide	-/-
X-11204	—	—	-/-
X-11244	lipid	lipid	-/-
X-11247	lipid	lipid	-/-
X-11255	amino acid	amino acid	-/-
X-11261	lipid	lipid	-/-
X-11299	—	—	-/-
X-11308	—	—	-/-
X-11315	amino acid	amino acid	-/-
X-11317	lipid	lipid	-/-
X-11319	lipid	lipid	-/-
X-11327	lipid	lipid	-/-

Table continued (p. 2 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-11333	amino acid	amino acid	-/-
X-11334	amino acid	amino acid	amino acid
X-11357	amino acid	-/-	amino acid
X-11372	—	—	-/-
X-11378	—	—	-/-
X-11381	lipid	lipid	-/-
X-11421	lipid	lipid	-/-
X-11422	nucleotide	nucleotide	-/-
X-11429	nucleotide	nucleotide	—
X-11437	—	—	-/-
X-11438	lipid	lipid	-/-
X-11440	lipid	lipid	lipid
X-11441	cofactors and vitamins	cofactors and vitamins	-/-
X-11442	cofactors and vitamins	cofactors and vitamins	-/-
X-11443	lipid	lipid	-/-
X-11444	—	—	—
X-11450	lipid	lipid	-/-
X-11452	xenobiotics	xenobiotics	—
X-11469	lipid	lipid	-/-
X-11470	lipid	lipid	lipid
X-11478	amino acid	amino acid	-/-
X-11483	—	—	-/-
X-11485	—	-/-	—
X-11491	lipid	lipid	-/-
X-11497	—	—	-/-
X-11508	lipid	lipid	-/-
X-11521	lipid	lipid	-/-
X-11522	cofactors and vitamins	cofactors and vitamins	-/-
X-11529	lipid	lipid	-/-
X-11530	cofactors and vitamins	cofactors and vitamins	-/-
X-11533	lipid	lipid	-/-
X-11537	lipid	lipid	-/-
X-11538	lipid	lipid	-/-
X-11540	lipid	lipid	lipid
X-11550	lipid	lipid	-/-
X-11552	lipid	lipid	-/-
X-11561	amino acid	peptide	-/-
X-11564	—	—	—
X-11612	nucleotide	-/-	nucleotide
X-11640	—	-/-	—
X-11787	amino acid	amino acid	amino acid
X-11792	peptide	peptide	-/-
X-11795	—	—	-/-
X-11799	lipid	lipid	-/-
X-11805	peptide	peptide	-/-

Table continued (p. 3 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-11809	cofactors and vitamins	cofactors and vitamins	-/-
X-11818	amino acid	amino acid	-/-
X-11820	lipid	lipid	-/-
X-11837	peptide	amino acid	-/-
X-11838	xenobiotics	xenobiotics	-/-
X-11843	—	—	—
X-11847	—	—	-/-
X-11849	—	—	-/-
X-11850	—	—	—
X-11852	—	-/-	—
X-11859	lipid	lipid	-/-
X-11880	—	—	-/-
X-11905	lipid	lipid	-/-
X-11979	—	-/-	—
X-12007	carbohydrate	carbohydrate	—
X-12013	—	-/-	—
X-12015	cofactors and vitamins	-/-	cofactors and vitamins
X-12026	amino acid	-/-	amino acid
X-12027	xenobiotics	-/-	xenobiotics
X-12029	—	—	-/-
X-12038	peptide	peptide	-/-
X-12039	xenobiotics	xenobiotics	-/-
X-12040	xenobiotics	xenobiotics	-/-
X-12051	lipid	lipid	-/-
X-12056	carbohydrate	carbohydrate	-/-
X-12063	lipid	lipid	-/-
X-12092	xenobiotics	—	-/-
X-12093	amino acid	amino acid	amino acid
X-12094	nucleotide	nucleotide	-/-
X-12096	amino acid	-/-	amino acid
X-12097	amino acid	-/-	amino acid
X-12100	cofactors and vitamins	-/-	cofactors and vitamins
X-12101	—	-/-	—
X-12111	amino acid	-/-	amino acid
X-12112	xenobiotics	-/-	xenobiotics
X-12115	amino acid	-/-	amino acid
X-12116	xenobiotics	cofactors and vitamins	-/-
X-12117	amino acid	-/-	amino acid
X-12119	amino acid	-/-	amino acid
X-12123	xenobiotics	-/-	xenobiotics
X-12124	amino acid	-/-	amino acid
X-12125	amino acid	-/-	amino acid
X-12126	—	-/-	—
X-12127	—	-/-	—
X-12170	amino acid	-/-	amino acid

Table continued (p. 4 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-12193	—	-/-	—
X-12199	nucleotide	-/-	nucleotide
X-12206	cofactors and vitamins	cofactors and vitamins	—
X-12212	xenobiotics	—	xenobiotics
X-12214	—	-/-	—
X-12216	peptide	amino acid	—
X-12217	xenobiotics	xenobiotics	-/-
X-12221	amino acid	-/-	amino acid
X-12230	xenobiotics	xenobiotics	—
X-12231	xenobiotics	xenobiotics	—
X-12236	—	—	-/-
X-12253	—	—	-/-
X-12257	amino acid	-/-	amino acid
X-12258	—	—	-/-
X-12261	—	-/-	—
X-12263	xenobiotics	-/-	xenobiotics
X-12267	xenobiotics	-/-	xenobiotics
X-12283	lipid	-/-	amino acid
X-12329	xenobiotics	xenobiotics	—
X-12379	amino acid	-/-	amino acid
X-12398	xenobiotics	-/-	xenobiotics
X-12405	amino acid	amino acid	-/-
X-12407	xenobiotics	xenobiotics	—
X-12410	xenobiotics	-/-	xenobiotics
X-12425	—	-/-	—
X-12450	lipid	lipid	-/-
X-12456	lipid	lipid	-/-
X-12465	lipid	lipid	-/-
X-12472	—	-/-	—
X-12511	amino acid	amino acid	amino acid
X-12524	—	—	-/-
X-12543	xenobiotics	amino acid	xenobiotics
X-12544	lipid	lipid	-/-
X-12556	amino acid	amino acid	-/-
X-12565	xenobiotics	-/-	xenobiotics
X-12627	lipid	lipid	-/-
X-12636	lipid	-/-	lipid
X-12644	lipid	lipid	-/-
X-12680	amino acid	-/-	amino acid
X-12686	amino acid	-/-	amino acid
X-12687	xenobiotics	-/-	xenobiotics
X-12688	amino acid	-/-	amino acid
X-12695	amino acid	-/-	amino acid
X-12696	carbohydrate	carbohydrate	-/-
X-12701	—	-/-	—

Table continued (p. 5 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-12704	xenobiotics	amino acid	amino acid
X-12706	—	-/-	—
X-12707	amino acid	-/-	amino acid
X-12708	amino acid	-/-	amino acid
X-12712	—	-/-	—
X-12713	amino acid	-/-	amino acid
X-12714	xenobiotics	-/-	xenobiotics
X-12718	xenobiotics	-/-	xenobiotics
X-12722	—	-/-	—
X-12726	amino acid	-/-	amino acid
X-12729	—	-/-	—
X-12730	xenobiotics	xenobiotics	xenobiotics
X-12731	—	-/-	—
X-12733	—	-/-	—
X-12734	amino acid	amino acid	-/-
X-12738	xenobiotics	-/-	xenobiotics
X-12739	—	-/-	—
X-12740	—	-/-	—
X-12742	—	—	-/-
X-12745	amino acid	-/-	amino acid
X-12749	xenobiotics	amino acid	-/-
X-12753	xenobiotics	-/-	xenobiotics
X-12776	—	—	-/-
X-12786	amino acid	amino acid	-/-
X-12798	lipid	lipid	-/-
X-12802	amino acid	amino acid	-/-
X-12811	—	-/-	—
X-12812	—	-/-	—
X-12814	—	-/-	—
X-12816	xenobiotics	xenobiotics	-/-
X-12818	xenobiotics	-/-	xenobiotics
X-12819	—	-/-	—
X-12820	—	-/-	—
X-12821	carbohydrate	-/-	carbohydrate
X-12822	lipid	—	lipid
X-12830	xenobiotics	xenobiotics	—
X-12831	—	-/-	—
X-12832	—	-/-	—
X-12834	—	-/-	—
X-12836	xenobiotics	-/-	xenobiotics
X-12839	lipid	-/-	amino acid
X-12840	—	-/-	—
X-12844	nucleotide	nucleotide	—
X-12845	—	-/-	—
X-12846	lipid	—	lipid

Table continued (p. 6 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-12847	xenobiotics	xenobiotics	xenobiotics
X-12849	—	—	—
X-12850	lipid	lipid	-/-
X-12851	—	—	-/-
X-12855	—	—	-/-
X-12860	lipid	—	lipid
X-12879	cofactors and vitamins	-/-	cofactors and vitamins
X-12906	—	-/-	—
X-13069	lipid	lipid	-/-
X-13183	—	—	-/-
X-13215	—	—	-/-
X-13255	—	-/-	—
X-13429	lipid	lipid	-/-
X-13430	—	-/-	—
X-13431	lipid	-/-	—
X-13435	lipid	lipid	-/-
X-13477	amino acid	amino acid	-/-
X-13507	—	-/-	—
X-13529	—	-/-	—
X-13548	—	—	-/-
X-13549	—	—	-/-
X-13553	amino acid	-/-	amino acid
X-13619	amino acid	amino acid	-/-
X-13671	—	—	-/-
X-13684	amino acid	-/-	amino acid
X-13688	—	-/-	—
X-13693	—	-/-	—
X-13698	amino acid	-/-	amino acid
X-13703	—	-/-	—
X-13722	—	-/-	—
X-13723	xenobiotics	-/-	xenobiotics
X-13726	xenobiotics	-/-	xenobiotics
X-13727	carbohydrate	carbohydrate	-/-
X-13728	xenobiotics	xenobiotics	xenobiotics
X-13729	amino acid	-/-	amino acid
X-13737	—	-/-	—
X-13834	lipid	-/-	lipid
X-13835	amino acid	amino acid	amino acid
X-13838	—	-/-	—
X-13844	amino acid	-/-	amino acid
X-13846	—	-/-	—
X-13847	—	-/-	—
X-13848	carbohydrate	amino acid	-/-
X-13859	—	—	-/-
X-13866	lipid	lipid	lipid

Table continued (p. 7 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-13874	lipid	-/-	lipid
X-13879	lipid	-/-	lipid
X-13891	lipid	lipid	-/-
X-14056	amino acid	cofactors and vitamins	amino acid
X-14057	—	—	-/-
X-14082	cofactors and vitamins	-/-	cofactors and vitamins
X-14086	peptide	peptide	-/-
X-14095	amino acid	amino acid	xenobiotics
X-14096	—	-/-	—
X-14099	amino acid	-/-	amino acid
X-14192	peptide	peptide	-/-
X-14196	amino acid	-/-	amino acid
X-14272	peptide	peptide	-/-
X-14302	amino acid	peptide	—
X-14314	amino acid	—	amino acid
X-14352	amino acid	amino acid	-/-
X-14364	—	—	—
X-14374	xenobiotics	xenobiotics	-/-
X-14473	peptide	peptide	-/-
X-14588	amino acid	amino acid	-/-
X-14625	amino acid	amino acid	-/-
X-14626	lipid	lipid	lipid
X-14632	lipid	lipid	-/-
X-14658	lipid	lipid	-/-
X-14662	lipid	lipid	lipid
X-14838	amino acid	-/-	amino acid
X-14939	lipid	lipid	-/-
X-14977	—	—	-/-
X-15136	—	-/-	—
X-15222	lipid	lipid	-/-
X-15245	carbohydrate	carbohydrate	-/-
X-15382	—	—	-/-
X-15461	amino acid	amino acid	amino acid
X-15469	lipid	-/-	lipid
X-15484	—	—	-/-
X-15486	lipid	—	lipid
X-15492	energy	lipid	energy
X-15497	lipid	xenobiotics	lipid
X-15503	amino acid	-/-	amino acid
X-15636	—	—	-/-
X-15646	lipid	-/-	lipid
X-15664	—	—	-/-
X-15666	amino acid	-/-	amino acid
X-15667	amino acid	amino acid	-/-
X-15728	xenobiotics	xenobiotics	xenobiotics

Table continued (p. 8 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-15754	—	-/-	—
X-15843	—	-/-	—
X-16071	lipid	amino acid	lipid
X-16083	—	-/-	—
X-16087	lipid	-/-	lipid
X-16123	peptide	peptide	-/-
X-16124	cofactors and vitamins	cofactors and vitamins	-/-
X-16125	—	—	-/-
X-16128	peptide	peptide	-/-
X-16130	peptide	peptide	-/-
X-16132	peptide	peptide	-/-
X-16134	peptide	peptide	-/-
X-16135	peptide	peptide	-/-
X-16136	—	—	-/-
X-16137	peptide	peptide	-/-
X-16206	—	—	-/-
X-16394	—	—	-/-
X-16397	amino acid	-/-	amino acid
X-16480	lipid	lipid	-/-
X-16563	amino acid	-/-	lipid
X-16564	xenobiotics	xenobiotics	-/-
X-16567	amino acid	-/-	lipid
X-16570	—	-/-	—
X-16578	lipid	lipid	-/-
X-16580	—	-/-	—
X-16654	lipid	lipid	lipid
X-16674	lipid	lipid	-/-
X-16774	nucleotide	-/-	nucleotide
X-16932	—	—	-/-
X-16934	—	—	-/-
X-16935	—	—	-/-
X-16940	xenobiotics	xenobiotics	-/-
X-16943	lipid	lipid	-/-
X-16946	cofactors and vitamins	cofactors and vitamins	-/-
X-16947	lipid	lipid	—
X-17010	—	-/-	—
X-17137	—	—	-/-
X-17138	amino acid	amino acid	-/-
X-17145	amino acid	amino acid	-/-
X-17150	xenobiotics	xenobiotics	-/-
X-17162	cofactors and vitamins	cofactors and vitamins	—
X-17174	peptide	peptide	-/-
X-17175	—	—	-/-
X-17179	—	—	-/-
X-17185	xenobiotics	xenobiotics	—

Table continued (p. 9 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-17189	peptide	peptide	-/-
X-17254	lipid	lipid	-/-
X-17269	lipid	lipid	-/-
X-17299	lipid	lipid	—
X-17301	—	-/-	—
X-17304	—	-/-	—
X-17314	xenobiotics	xenobiotics	-/-
X-17323	amino acid	-/-	cofactors and vitamins
X-17324	lipid	-/-	lipid
X-17325	peptide	-/-	peptide
X-17327	—	-/-	—
X-17328	—	-/-	—
X-17335	—	-/-	—
X-17336	—	—	-/-
X-17337	lipid	-/-	lipid
X-17339	—	-/-	—
X-17340	lipid	-/-	lipid
X-17342	xenobiotics	-/-	xenobiotics
X-17343	—	-/-	—
X-17346	—	-/-	—
X-17349	—	-/-	—
X-17350	—	-/-	—
X-17351	amino acid	-/-	amino acid
X-17353	—	-/-	—
X-17354	amino acid	-/-	amino acid
X-17357	—	—	—
X-17358	—	-/-	—
X-17359	—	—	—
X-17361	—	-/-	—
X-17365	amino acid	-/-	amino acid
X-17367	xenobiotics	-/-	xenobiotics
X-17370	—	-/-	—
X-17371	—	-/-	—
X-17393	xenobiotics	-/-	xenobiotics
X-17398	—	-/-	—
X-17438	lipid	lipid	—
X-17441	peptide	peptide	-/-
X-17443	lipid	lipid	-/-
X-17502	carbohydrate	carbohydrate	-/-
X-17612	cofactors and vitamins	cofactors and vitamins	-/-
X-17654	—	—	-/-
X-17673	xenobiotics	-/-	xenobiotics
X-17674	xenobiotics	-/-	xenobiotics
X-17675	—	-/-	—
X-17677	xenobiotics	-/-	xenobiotics

Table continued (p. 10 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-17682	—	-/-	—
X-17685	xenobiotics	amino acid	xenobiotics
X-17688	amino acid	-/-	amino acid
X-17689	xenobiotics	-/-	xenobiotics
X-17690	xenobiotics	-/-	xenobiotics
X-17692	—	-/-	—
X-17693	—	-/-	—
X-17699	lipid	-/-	lipid
X-17701	—	-/-	—
X-17703	—	—	-/-
X-17704	—	-/-	—
X-17705	lipid	-/-	—
X-17706	—	—	-/-
X-17723	—	-/-	—
X-17735	—	-/-	—
X-17765	lipid	-/-	lipid
X-17807	—	-/-	—
X-18059	—	-/-	—
X-18221	carbohydrate	carbohydrate	-/-
X-18240	xenobiotics	-/-	xenobiotics
X-18249	—	—	-/-
X-18345	—	-/-	—
X-18601	lipid	lipid	-/-
X-18603	amino acid	-/-	amino acid
X-18606	—	-/-	—
X-18752	—	—	-/-
X-18774	—	—	-/-
X-18779	amino acid	-/-	amino acid
X-18838	—	-/-	—
X-18886	—	-/-	—
X-18887	amino acid	-/-	amino acid
X-18888	lipid	-/-	lipid
X-18889	amino acid	-/-	amino acid
X-18935	—	-/-	—
X-18938	lipid	-/-	lipid
X-19141	lipid	-/-	lipid
X-19299	—	-/-	—
X-19434	lipid	-/-	lipid
X-19497	—	-/-	—
X-19561	lipid	-/-	lipid
X-19913	nucleotide	-/-	nucleotide
X-19932	—	-/-	—
X-20624	peptide	-/-	peptide
X-21258	amino acid	-/-	amino acid
X-21283	nucleotide	-/-	nucleotide

Table continued (p. 11 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-21295	—	-/-	—
X-21410	lipid	-/-	lipid
X-21448	—	-/-	—
X-21467	lipid	-/-	lipid
X-21470	lipid	-/-	lipid
X-21474	—	-/-	—
X-21659	—	-/-	—
X-21668	lipid	-/-	lipid
X-21736	—	-/-	—
X-21737	xenobiotics	-/-	xenobiotics
X-21785	amino acid	-/-	amino acid
X-21788	lipid	-/-	amino acid
X-21792	lipid	-/-	lipid
X-21796	—	-/-	—
X-21803	—	-/-	—
X-21804	xenobiotics	-/-	xenobiotics
X-21807	—	-/-	—
X-21815	xenobiotics	-/-	xenobiotics
X-21821	lipid	-/-	amino acid
X-21825	—	-/-	—
X-21826	—	-/-	—
X-21827	lipid	-/-	lipid
X-21828	—	-/-	—
X-21829	lipid	-/-	lipid
X-21830	—	-/-	—
X-21831	lipid	-/-	lipid
X-21834	amino acid	-/-	amino acid
X-21835	—	-/-	—
X-21839	—	-/-	—
X-21840	—	-/-	—
X-21841	—	-/-	—
X-21842	amino acid	-/-	amino acid
X-21844	—	-/-	—
X-21845	—	-/-	—
X-21846	lipid	-/-	—
X-21849	—	-/-	—
X-21850	—	-/-	—
X-21851	lipid	-/-	lipid
X-22035	—	-/-	—
X-22102	—	-/-	—
X-22143	cofactors and vitamins	-/-	cofactors and vitamins
X-22147	—	-/-	—
X-22158	lipid	-/-	lipid
X-22379	lipid	-/-	lipid
X-22475	amino acid	-/-	amino acid

Table continued (p. 12 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-22508	xenobiotics	-/-	xenobiotics
X-22509	—	-/-	—
X-22515	—	-/-	—
X-22522	—	-/-	—
X-22755	—	-/-	—
X-22757	lipid	-/-	lipid
X-22834	lipid	-/-	lipid
X-22836	amino acid	-/-	amino acid
X-23039	—	-/-	—
X-23157	—	-/-	—
X-23161	—	-/-	—
X-23164	cofactors and vitamins	-/-	cofactors and vitamins
X-23193	—	-/-	—
X-23196	lipid	-/-	lipid
X-23291	xenobiotics	-/-	xenobiotics
X-23299	—	-/-	—
X-23304	xenobiotics	-/-	xenobiotics
X-23306	xenobiotics	-/-	xenobiotics
X-23308	—	-/-	—
X-23311	—	-/-	—
X-23314	carbohydrate	-/-	carbohydrate
X-23329	—	-/-	—
X-23331	—	-/-	—
X-23369	—	-/-	—
X-23423	amino acid	-/-	amino acid
X-23429	xenobiotics	-/-	xenobiotics
X-23438	—	-/-	—
X-23445	—	-/-	—
X-23461	—	-/-	—
X-23512	—	-/-	—
X-23518	amino acid	-/-	amino acid
X-23581	amino acid	-/-	amino acid
X-23583	amino acid	-/-	amino acid
X-23584	carbohydrate	-/-	carbohydrate
X-23590	amino acid	-/-	amino acid
X-23593	peptide	-/-	peptide
X-23641	—	-/-	—
X-23644	xenobiotics	-/-	xenobiotics
X-23645	amino acid	-/-	amino acid
X-23647	amino acid	-/-	amino acid
X-23648	amino acid	-/-	amino acid
X-23649	xenobiotics	-/-	xenobiotics
X-23650	—	-/-	—
X-23652	—	-/-	—
X-23653	amino acid	-/-	amino acid

Table continued (p. 13 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-23654	—	-/-	—
X-23655	xenobiotics	-/-	xenobiotics
X-23656	amino acid	-/-	amino acid
X-23657	—	-/-	—
X-23659	xenobiotics	-/-	xenobiotics
X-23662	amino acid	-/-	amino acid
X-23666	—	-/-	—
X-23668	amino acid	-/-	amino acid
X-23678	—	-/-	—
X-23680	—	-/-	—
X-23729	—	-/-	—
X-23739	amino acid	-/-	amino acid
X-23744	—	-/-	—
X-23747	amino acid	-/-	amino acid
X-23748	—	-/-	—
X-23776	xenobiotics	-/-	xenobiotics
X-23780	amino acid	-/-	amino acid
X-23908	amino acid	-/-	amino acid
X-23997	amino acid	-/-	amino acid
X-24228	carbohydrate	-/-	carbohydrate
X-24249	amino acid	-/-	amino acid
X-24270	xenobiotics	-/-	xenobiotics
X-24272	—	-/-	—
X-24293	xenobiotics	-/-	xenobiotics
X-24327	—	-/-	—
X-24328	lipid	-/-	lipid
X-24329	amino acid	-/-	amino acid
X-24330	lipid	-/-	lipid
X-24332	—	-/-	—
X-24333	—	-/-	—
X-24334	—	-/-	—
X-24337	—	-/-	—
X-24338	amino acid	-/-	amino acid
X-24339	lipid	-/-	lipid
X-24340	—	-/-	—
X-24341	—	-/-	—
X-24342	—	-/-	—
X-24343	xenobiotics	-/-	xenobiotics
X-24344	xenobiotics	-/-	xenobiotics
X-24345	peptide	-/-	peptide
X-24346	—	-/-	—
X-24347	lipid	-/-	amino acid
X-24348	lipid	-/-	—
X-24349	lipid	-/-	lipid
X-24350	peptide	-/-	peptide

Table continued (p. 14 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-24352	—	-/-	—
X-24353	amino acid	-/-	amino acid
X-24354	—	-/-	—
X-24357	lipid	-/-	lipid
X-24358	lipid	-/-	lipid
X-24359	cofactors and vitamins	-/-	cofactors and vitamins
X-24360	lipid	-/-	lipid
X-24361	lipid	-/-	lipid
X-24362	lipid	-/-	lipid
X-24363	—	-/-	—
X-24387	amino acid	-/-	amino acid
X-24400	cofactors and vitamins	-/-	cofactors and vitamins
X-24402	amino acid	-/-	amino acid
X-24403	—	-/-	—
X-24406	—	-/-	—
X-24408	nucleotide	-/-	nucleotide
X-24410	amino acid	-/-	amino acid
X-24411	amino acid	-/-	amino acid
X-24412	xenobiotics	-/-	xenobiotics
X-24414	xenobiotics	-/-	xenobiotics
X-24415	—	-/-	—
X-24416	cofactors and vitamins	-/-	cofactors and vitamins
X-24417	lipid	-/-	lipid
X-24418	—	-/-	—
X-24419	—	-/-	—
X-24422	cofactors and vitamins	-/-	cofactors and vitamins
X-24431	cofactors and vitamins	-/-	cofactors and vitamins
X-24432	cofactors and vitamins	-/-	cofactors and vitamins
X-24452	amino acid	-/-	amino acid
X-24455	amino acid	-/-	amino acid
X-24456	—	-/-	—
X-24460	energy	-/-	energy
X-24462	carbohydrate	-/-	carbohydrate
X-24465	amino acid	-/-	amino acid
X-24466	amino acid	-/-	amino acid
X-24468	amino acid	-/-	amino acid
X-24473	xenobiotics	-/-	xenobiotics
X-24475	xenobiotics	-/-	xenobiotics
X-24490	amino acid	-/-	amino acid
X-24494	lipid	-/-	lipid
X-24497	—	-/-	—
X-24498	xenobiotics	-/-	xenobiotics
X-24512	amino acid	-/-	amino acid
X-24513	amino acid	-/-	amino acid
X-24514	amino acid	-/-	amino acid

Table continued (p. 15 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-24515	—	-/-	—
X-24518	amino acid	-/-	amino acid
X-24519	amino acid	-/-	amino acid
X-24520	—	-/-	—
X-24522	lipid	-/-	lipid
X-24527	—	-/-	—
X-24528	—	-/-	—
X-24542	amino acid	-/-	amino acid
X-24543	xenobiotics	-/-	xenobiotics
X-24546	lipid	-/-	lipid
X-24554	—	-/-	—
X-24556	—	-/-	—
X-24557	—	-/-	—
X-24586	amino acid	-/-	amino acid
X-24587	amino acid	-/-	amino acid
X-24588	lipid	-/-	lipid
X-24590	amino acid	-/-	amino acid
X-24623	—	-/-	—
X-24660	lipid	-/-	lipid
X-24669	lipid	-/-	lipid
X-24680	—	-/-	—
X-24682	—	-/-	—
X-24686	carbohydrate	-/-	carbohydrate
X-24699	—	-/-	—
X-24736	amino acid	-/-	amino acid
X-24738	lipid	-/-	lipid
X-24757	xenobiotics	-/-	xenobiotics
X-24758	amino acid	-/-	amino acid
X-24759	—	-/-	—
X-24760	xenobiotics	-/-	xenobiotics
X-24762	—	-/-	—
X-24764	—	-/-	—
X-24766	amino acid	-/-	amino acid
X-24768	—	-/-	—
X-24795	xenobiotics	-/-	xenobiotics
X-24796	amino acid	-/-	amino acid
X-24798	lipid	-/-	—
X-24799	lipid	-/-	lipid
X-24801	—	-/-	—
X-24807	lipid	-/-	lipid
X-24808	amino acid	-/-	amino acid
X-24809	amino acid	-/-	amino acid
X-24811	xenobiotics	-/-	xenobiotics
X-24812	—	-/-	—
X-24834	peptide	-/-	—

Table continued (p. 16 / 16)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-24838	lipid	-/-	lipid
X-24860	lipid	-/-	lipid
X-24969	amino acid	-/-	amino acid
X-24971	amino acid	-/-	amino acid
X-24983	xenobiotics	-/-	xenobiotics

Table B.5: Predicted super pathways of unknown metabolites in SUMMIT/GCKD, SUMMIT, and GCKD

Super pathways were predicted with three different network models based on: SUMMIT/GCKD, SUMMIT, or GCKD. A subset of 59 unknown metabolites was measured in SUMMIT as well as in GCKD by the established platform and by the new high-resolution platform of Metabolon^{TM}, respectively. For 427 of 677 unknown metabolites, super pathways were predicted in at least two of three sets. '-/-' and '—' indicate that the unknown metabolite was not measured in the respective cohort, or that the unknown was measured, but no super pathway was predicted, respectively. The table shows all unknown metabolites measured in SUMMIT or GCKD, even if there is no super pathway prediction.

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-01911	tyrosine metabolism	ascorbate and aldarate metabolism	—
X-02249	essential fatty acid, or fatty acid, dicarboxylate, or long chain fatty acid	essential fatty acid, or fatty acid, dicarboxylate, or long chain fatty acid	tryptophan metabolism
X-02269	fatty acid, dicarboxylate, or long chain fatty acid	fatty acid, dicarboxylate, or long chain fatty acid	-/-
X-02973	ascorbate and aldarate metabolism	ascorbate and aldarate metabolism	-/-
X-03088	urea cycle; arginine and proline metabolism	urea cycle; arginine-, proline-, metabolism	-/-
X-03094	sterol/steroid	sterol/steroid	-/-
X-04357	fructose, mannose, galactose, starch, and sucrose metabolism	fructose, mannose, galactose, starch, and sucrose metabolism	-/-
X-04495	glutathione metabolism	butanoate metabolism, or creatine metabolism, or cysteine, methionine, sam, taurine metabolism	-/-
X-05491	butanoate metabolism	butanoate metabolism	-/-
X-06126	benzoate metabolism	phenylalanine & tyrosine metabolism	-/-
X-06227	oxidative phosphorylation	—	-/-
X-06246	alanine and aspartate metabolism	alanine and aspartate metabolism	-/-
X-06267	urea cycle; arginine and proline metabolism	urea cycle; arginine-, proline-, metabolism	-/-
X-06307	dipeptide	dipeptide	-/-
X-08402	sphingolipid metabolism, or sterol/steroid	sphingolipid, or sterol/steroid	-/-
X-08988	leucine, isoleucine and valine metabolism	glycine, serine and threonine metabolism	-/-
X-09706	urea cycle; arginine and proline metabolism, or valine, leucine and isoleucine metabolism	urea cycle; arginine-, proline-, metabolism, or valine, leucine and isoleucine metabolism	-/-
X-09789	chemical	—	chemical
X-10419	sterol/steroid	sterol/steroid	-/-
X-10500	sterol/steroid	sterol/steroid	-/-
X-10506	alanine and aspartate metabolism	alanine and aspartate metabolism	-/-
X-10510	sphingolipid metabolism, or sterol/steroid	sphingolipid, or sterol/steroid	-/-
X-10810	purine metabolism, (hypo)xanthine/inosine containing	purine metabolism, (hypo)xanthine/inosine containing	-/-
X-11244	androgenic steroids	sterol, steroid	-/-
X-11247	long chain fatty acid	long chain fatty acid	-/-
X-11255	leucine, isoleucine and valine metabolism	valine, leucine and isoleucine metabolism	-/-
X-11261	fatty acid metabolism(acyl carnitine)	carnitine metabolism	-/-
X-11315	glutamate metabolism	glutamate metabolism	-/-
X-11317	lysolipid	lysolipid	-/-
X-11319	long chain fatty acid	long chain fatty acid	-/-
X-11327	fatty acid metabolism(acyl carnitine)	carnitine metabolism	-/-

Table continued (p. 2 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-11333	amino fatty acid, or lysine metabolism, or urea cycle; arginine and proline metabolism	amino fatty acid, or lysine metabolism, or urea cycle; arginine-, proline-, metabolism	-/-
X-11334	histidine metabolism, or lysine metabolism	lysine metabolism	histidine metabolism
X-11357	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-11381	fatty acid metabolism (also bcaa metabolism), or fatty acid metabolism(acyl carnitine)	carnitine metabolism	-/-
X-11421	fatty acid metabolism(acyl carnitine)	carnitine metabolism	-/-
X-11422	purine metabolism, (hypo)xanthine/inosine containing	purine metabolism, (hypo)xanthine/inosine containing	-/-
X-11429	purine metabolism, (hypo)xanthine/inosine containing, or pyrimidine metabolism, uracil containing	purine metabolism, (hypo)xanthine/inosine containing, or pyrimidine metabolism, uracil containing	—
X-11438	fatty acid, dicarboxylate	fatty acid, dicarboxylate, or long chain fatty acid	-/-
X-11440	androgenic steroids	sterol, steroid	androgenic steroids
X-11441	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-11442	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-11443	androgenic steroids	sterol, steroid	-/-
X-11450	androgenic steroids	sterol, steroid	-/-
X-11452	food component/plant	food component/plant	—
X-11469	fatty acid metabolism (acyl glutamine), or fatty acid, dicarboxylate, or long chain fatty acid	sterol, steroid	-/-
X-11470	progestin steroids	sterol/steroid	progestin steroids
X-11478	tryptophan metabolism	phenylalanine & tyrosine metabolism, or tryptophan metabolism	-/-
X-11491	fatty acid, dicarboxylate, or lysolipid	fatty acid, dicarboxylate, or lysolipid	-/-
X-11508	fatty acid, monohydroxy	fatty acid, monohydroxy	-/-
X-11521	essential fatty acid	essential fatty acid	-/-
X-11522	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-11529	fatty acid, dicarboxylate, or lysolipid	fatty acid, dicarboxylate, or lysolipid	-/-
X-11530	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-11533	medium chain fatty acid	medium chain fatty acid	-/-
X-11537	phospholipid metabolism	glycerolipid metabolism	-/-
X-11538	fatty acid, dicarboxylate	fatty acid, dicarboxylate	-/-
X-11540	fatty acid metabolism(acyl carnitine), or phospholipid metabolism	glycerolipid metabolism	fatty acid metabolism(acyl carnitine)
X-11550	medium chain fatty acid	medium chain fatty acid	-/-
X-11552	fatty acid, amide	fatty acid, amide	-/-
X-11561	alanine and aspartate metabolism	dipeptide	-/-

Table continued (p. 3 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-11612	purine metabolism, (hypo)xanthine/inosine containing	-/-	purine metabolism, (hypo)xanthine/inosine containing
X-11787	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	guanidino and acetamido metabolism
X-11792	dipeptide	dipeptide	-/-
X-11799	inositol metabolism	inositol metabolism	-/-
X-11805	dipeptide	dipeptide	-/-
X-11809	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-11818	amino fatty acid	amino fatty acid	-/-
X-11820	carnitine metabolism, or sterol/steroid	carnitine metabolism, or sterol/steroid	-/-
X-11837	acetylated peptides	phenylalanine & tyrosine metabolism	-/-
X-11838	drug	drug	-/-
X-11859	medium chain fatty acid	medium chain fatty acid	-/-
X-11905	fatty acid, dicarboxylate	fatty acid, dicarboxylate	-/-
X-12007	disaccharides and oligosaccharides	fructose, mannose, galactose, starch, and sucrose metabolism	—
X-12015	pantothenate and coa metabolism	-/-	pantothenate and coa metabolism
X-12026	histidine metabolism	-/-	histidine metabolism
X-12027	food component/plant	-/-	food component/plant
X-12038	polypeptide	polypeptide	-/-
X-12039	food component/plant, or xanthine metabolism	food component/plant, or xanthine metabolism	-/-
X-12040	food component/plant	food component, plant	-/-
X-12051	lysolipid	lysolipid	-/-
X-12056	fructose, mannose, galactose, starch, and sucrose metabolism	fructose, mannose, galactose, starch, and sucrose metabolism	-/-
X-12063	androgenic steroids	sterol, steroid	-/-
X-12092	chemical	—	-/-
X-12093	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	histidine metabolism, or urea cycle; arginine and proline metabolism
X-12094	nad metabolism	nad metabolism	-/-
X-12096	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism
X-12097	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism
X-12100	nicotinate and nicotinamide metabolism, or pterin metabolism	-/-	nicotinate and nicotinamide metabolism, or pterin metabolism
X-12111	glutamate metabolism, or urea cycle; arginine and proline metabolism	-/-	glutamate metabolism, or urea cycle; arginine and proline metabolism
X-12112	chemical	-/-	chemical
X-12115	glutamate metabolism, or leucine, isoleucine and valine metabolism, or polyamine metabolism	-/-	glutamate metabolism, or leucine, isoleucine and valine metabolism, or polyamine metabolism

Table continued (p. 4 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-12116	food component/plant	ascorbate and aldarate metabolism	-/-
X-12117	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism
X-12119	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-12123	food component/plant	-/-	food component/plant
X-12124	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-12125	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-12170	tryptophan metabolism	-/-	tryptophan metabolism
X-12199	purine metabolism, guanine containing	-/-	purine metabolism, guanine containing
X-12206	ascorbate and aldarate metabolism	ascorbate and aldarate metabolism	—
X-12212	food component/plant	—	food component/plant
X-12216	acetylated peptides	phenylalanine & tyrosine metabolism	—
X-12217	benzoate metabolism	benzoate metabolism	-/-
X-12221	tyrosine metabolism	-/-	tyrosine metabolism
X-12230	benzoate metabolism	benzoate metabolism	—
X-12231	food component/plant	food component/plant	—
X-12257	tyrosine metabolism	-/-	tyrosine metabolism
X-12263	food component/plant	-/-	food component/plant
X-12267	food component/plant	-/-	food component/plant
X-12283	long chain fatty acid	-/-	tryptophan metabolism
X-12329	food component/plant	food component/plant	—
X-12379	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-12398	drug	-/-	drug
X-12405	tryptophan metabolism	tryptophan metabolism	-/-
X-12407	food component/plant	food component/plant	—
X-12410	xanthine metabolism	-/-	xanthine metabolism
X-12450	essential fatty acid, or fatty acid, monohydroxy	essential fatty acid, or fatty acid, monohydroxy	-/-
X-12456	androgenic steroids	sterol, steroid	-/-
X-12465	fatty acid metabolism(acyl carnitine), or ketone bodies	carnitine metabolism, or ketone bodies	-/-
X-12511	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	urea cycle; arginine and proline metabolism
X-12543	benzoate metabolism	phenylalanine & tyrosine metabolism	benzoate metabolism
X-12544	androgenic steroids	sterol/steroid	-/-
X-12556	glycine, serine and threonine metabolism	glycine, serine and threonine metabolism	-/-
X-12565	food component/plant	-/-	food component/plant
X-12627	essential fatty acid, or long chain fatty acid	essential fatty acid, or long chain fatty acid	-/-

Table continued (p. 5 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-12636	fatty acid metabolism(acyl carnitine)	-/-	fatty acid metabolism (acyl glutamine)
X-12644	lysolipid	lysolipid	-/-
X-12680	lysine metabolism	-/-	lysine metabolism
X-12686	creatine metabolism	-/-	creatine metabolism
X-12687	chemical	-/-	chemical
X-12688	polyamine metabolism	-/-	histidine metabolism, or polyamine metabolism
X-12695	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-12696	glycolysis, gluconeogenesis, and pyruvate metabolism	glycolysis, gluconeogenesis, pyruvate metabolism	-/-
X-12704	benzoate metabolism	phenylalanine & tyrosine metabolism	—
X-12707	tyrosine metabolism	-/-	tyrosine metabolism
X-12708	tyrosine metabolism	-/-	tyrosine metabolism
X-12713	tyrosine metabolism	-/-	tyrosine metabolism
X-12714	chemical, or food component/plant	-/-	chemical, or food component/plant
X-12718	chemical	-/-	chemical
X-12726	tyrosine metabolism	-/-	tyrosine metabolism
X-12730	food component/plant	food component/plant	food component/plant
X-12734	tyrosine metabolism	phenylalanine & tyrosine metabolism	-/-
X-12738	benzoate metabolism	-/-	benzoate metabolism
X-12745	tryptophan metabolism	-/-	tryptophan metabolism
X-12749	chemical, or drug	phenylalanine & tyrosine metabolism	-/-
X-12753	food component/plant	-/-	food component/plant
X-12786	alanine and aspartate metabolism	alanine and aspartate metabolism	-/-
X-12798	fatty acid metabolism(acyl carnitine)	carnitine metabolism	-/-
X-12802	leucine, isoleucine and valine metabolism	valine, leucine and isoleucine metabolism	-/-
X-12816	food component/plant	food component/plant	-/-
X-12818	food component/plant	-/-	food component/plant
X-12821	pentose metabolism	-/-	pentose metabolism
X-12822	fatty acid metabolism (acyl glutamine)	—	fatty acid metabolism (acyl glutamine)
X-12830	food component/plant	food component/plant	—
X-12836	food component/plant	-/-	food component/plant
X-12839	fatty acid, dicarboxylate, or long chain fatty acid	-/-	—
X-12844	nad metabolism	nad metabolism	—
X-12846	androgenic steroids	—	androgenic steroids
X-12847	food component/plant	food component/plant	food component/plant
X-12850	androgenic steroids	sterol, steroid	-/-
X-12860	fatty acid metabolism(acyl carnitine)	—	fatty acid metabolism (acyl glutamine)
X-12879	pantothenate and coa metabolism, or vitamin b6 metabolism	-/-	pantothenate and coa metabolism, or vitamin b6 metabolism

Table continued (p. 6 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-13069	long chain fatty acid	long chain fatty acid	-/-
X-13429	androgenic steroids	sterol, steroid	-/-
X-13431	carnitine metabolism	-/-	—
X-13435	fatty acid metabolism(acyl carnitine)	carnitine metabolism	-/-
X-13477	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	-/-
X-13553	tyrosine metabolism	-/-	tyrosine metabolism
X-13619	urea cycle; arginine and proline metabolism	urea cycle; arginine-, proline-, metabolism	-/-
X-13684	phenylalanine metabolism	-/-	phenylalanine metabolism, or tyrosine metabolism
X-13698	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-13723	food component/plant	-/-	food component/plant
X-13726	benzoate metabolism	-/-	benzoate metabolism
X-13727	disaccharides and oligosaccharides	fructose, mannose, galactose, starch, and sucrose metabolism	-/-
X-13728	xanthine metabolism	xanthine metabolism	xanthine metabolism
X-13729	tryptophan metabolism, or urea cycle; arginine and proline metabolism	-/-	tryptophan metabolism, or urea cycle; arginine and proline metabolism
X-13834	fatty acid, dicarboxylate	-/-	fatty acid, dicarboxylate
X-13835	histidine metabolism	histidine metabolism	histidine metabolism
X-13844	histidine metabolism	-/-	histidine metabolism
X-13848	fructose, mannose, galactose, starch, and sucrose metabolism	—	-/-
X-13866	fatty acid metabolism (acyl glutamine)	long chain fatty acid	fatty acid metabolism (acyl glutamine)
X-13874	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-13879	fatty acid, dicarboxylate	-/-	fatty acid, dicarboxylate
X-13891	fatty acid, dicarboxylate	fatty acid, dicarboxylate	-/-
X-14056	methionine, cysteine, sam and taurine metabolism	hemoglobin and porphyrin metabolism	methionine, cysteine, sam and taurine metabolism
X-14082	nicotinate and nicotinamide metabolism	-/-	nicotinate and nicotinamide metabolism
X-14086	dipeptide	dipeptide	-/-
X-14095	—	—	xanthine metabolism
X-14099	glutamate metabolism	-/-	glutamate metabolism
X-14192	dipeptide	dipeptide	-/-
X-14196	leucine, isoleucine and valine metabolism	-/-	leucine, isoleucine and valine metabolism
X-14272	dipeptide, or gamma-glutamyl amino acid	dipeptide, or gamma-glutamyl amino acid	-/-
X-14302	tryptophan metabolism	polypeptide	—
X-14314	glutathione metabolism	—	glutathione metabolism
X-14352	leucine, isoleucine and valine metabolism, or urea cycle; arginine and proline metabolism	urea cycle; arginine-, proline-, metabolism, or valine, leucine and isoleucine metabolism	-/-
X-14374	benzoate metabolism, or xanthine metabolism	benzoate metabolism, or xanthine metabolism	-/-

Table continued (p. 7 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-14473	dipeptide	dipeptide	-/-
X-14588	lysine metabolism	lysine metabolism	-/-
X-14625	glutamate metabolism, or glutathione metabolism	glutamate metabolism, or glutathione metabolism	-/-
X-14626	fatty acid, dicarboxylate, or secondary bile acid metabolism	fatty acid, dicarboxylate, or lysolipid	secondary bile acid metabolism
X-14632	secondary bile acid metabolism, or sterol/steroid	bile acid metabolism, or sterol/steroid	-/-
X-14658	bile acid metabolism, or secondary bile acid metabolism	bile acid metabolism	-/-
X-14662	secondary bile acid metabolism	bile acid metabolism	secondary bile acid metabolism
X-14838	histidine metabolism	-/-	histidine metabolism
X-14939	medium chain fatty acid	medium chain fatty acid	-/-
X-15222	medium chain fatty acid	medium chain fatty acid	-/-
X-15245	glycolysis, gluconeogenesis, and pyruvate metabolism	glycolysis, gluconeogenesis, pyruvate metabolism	-/-
X-15461	lysine metabolism	phenylalanine & tyrosine metabolism	lysine metabolism
X-15469	fatty acid metabolism(acyl carnitine)	-/-	fatty acid metabolism(acyl carnitine)
X-15486	fatty acid metabolism(acyl glycine)	—	fatty acid metabolism(acyl glycine)
X-15492	tca cycle	sterol/steroid	tca cycle
X-15497	fatty acid metabolism (acyl glutamine)	drug	fatty acid metabolism (acyl glutamine)
X-15503	tryptophan metabolism	-/-	tryptophan metabolism
X-15646	fatty acid metabolism (acyl glutamine), or secondary bile acid metabolism	-/-	fatty acid metabolism (acyl glutamine), or secondary bile acid metabolism
X-15666	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-15667	tryptophan metabolism	tryptophan metabolism	-/-
X-15728	benzoate metabolism, or food component/plant	benzoate metabolism	food component/plant
X-16071	fatty acid metabolism (acyl glutamine)	tryptophan metabolism	fatty acid metabolism (acyl glutamine)
X-16087	long chain fatty acid	-/-	long chain fatty acid
X-16123	dipeptide	dipeptide	-/-
X-16124	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-16128	dipeptide	dipeptide	-/-
X-16130	dipeptide	dipeptide	-/-
X-16132	dipeptide	dipeptide	-/-
X-16134	fibrinogen cleavage peptide	fibrinogen cleavage peptide	-/-
X-16135	fibrinogen cleavage peptide	fibrinogen cleavage peptide	-/-
X-16137	polypeptide	polypeptide	-/-
X-16397	lysine metabolism	-/-	lysine metabolism
X-16480	essential fatty acid, or fatty acid, dicarboxylate	essential fatty acid, or fatty acid, dicarboxylate	-/-
X-16563	leucine, isoleucine and valine metabolism	-/-	fatty acid metabolism (acyl glutamine)

Table continued (p. 8 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-16564	benzoate metabolism	benzoate metabolism	-/-
X-16567	leucine, isoleucine and valine metabolism	-/-	fatty acid metabolism (acyl glutamine)
X-16578	medium chain fatty acid	medium chain fatty acid	-/-
X-16654	secondary bile acid metabolism	bile acid metabolism	secondary bile acid metabolism
X-16674	fatty acid, monohydroxy	fatty acid, monohydroxy	-/-
X-16774	purine metabolism, (hypo)xanthine/inosine containing	-/-	purine metabolism, (hypo)xanthine/inosine containing
X-16940	benzoate metabolism	benzoate metabolism	-/-
X-16943	medium chain fatty acid	medium chain fatty acid	-/-
X-16946	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-
X-16947	inositol metabolism	inositol metabolism	—
X-17138	leucine, isoleucine and valine metabolism	valine, leucine and isoleucine metabolism	-/-
X-17145	tryptophan metabolism	tryptophan metabolism	-/-
X-17150	food component/plant	food component/plant	-/-
X-17162	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	—
X-17174	dipeptide	dipeptide	-/-
X-17185	benzoate metabolism	benzoate metabolism	—
X-17189	dipeptide	dipeptide	-/-
X-17254	lysolipid	lysolipid	-/-
X-17269	medium chain fatty acid	medium chain fatty acid	-/-
X-17299	carnitine metabolism	carnitine metabolism	—
X-17314	benzoate metabolism	benzoate metabolism	-/-
X-17323	leucine, isoleucine and valine metabolism, or phenylalanine metabolism, or polyamine metabolism	-/-	nicotinate and nicotinamide metabolism
X-17324	long chain fatty acid	-/-	long chain fatty acid
X-17325	acetylated peptides	-/-	acetylated peptides
X-17337	carnitine metabolism	-/-	carnitine metabolism
X-17340	progestin steroids	-/-	progestin steroids
X-17342	benzoate metabolism	-/-	benzoate metabolism
X-17351	tryptophan metabolism	-/-	tryptophan metabolism
X-17354	methionine, cysteine, sam and taurine metabolism	-/-	methionine, cysteine, sam and taurine metabolism
X-17365	leucine, isoleucine and valine metabolism, or methionine, cysteine, sam and taurine metabolism	-/-	leucine, isoleucine and valine metabolism, or methionine, cysteine, sam and taurine metabolism
X-17367	benzoate metabolism	-/-	benzoate metabolism
X-17393	chemical, or food component/plant	-/-	chemical, or food component/plant
X-17438	fatty acid, dicarboxylate	fatty acid, dicarboxylate	—
X-17441	fibrinogen cleavage peptide	fibrinogen cleavage peptide	-/-
X-17443	fatty acid, monohydroxy	fatty acid, monohydroxy	-/-
X-17502	glycolysis, gluconeogenesis, and pyruvate metabolism	glycolysis, gluconeogenesis, pyruvate metabolism	-/-
X-17612	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	-/-

Table continued (p. 9 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-17673	chemical	-/-	chemical
X-17674	food component/plant	-/-	food component/plant
X-17677	food component/plant	-/-	food component/plant
X-17685	benzoate metabolism	phenylalanine & tyrosine metabolism	benzoate metabolism
X-17688	tryptophan metabolism	-/-	tryptophan metabolism
X-17689	benzoate metabolism	-/-	benzoate metabolism
X-17690	food component/plant	-/-	food component/plant
X-17699	fatty acid metabolism(acyl glycine)	-/-	fatty acid metabolism(acyl glycine)
X-17705	sterol/steroid	-/-	—
X-17765	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-18221	glycolysis, gluconeogenesis, and pyruvate metabolism	glycolysis, gluconeogenesis, pyruvate metabolism	-/-
X-18240	food component/plant	-/-	food component/plant
X-18601	androgenic steroids	sterol, steroid	-/-
X-18603	glycine, serine and threonine metabolism	-/-	glycine, serine and threonine metabolism
X-18779	tyrosine metabolism	-/-	tyrosine metabolism
X-18887	lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	lysine metabolism, or urea cycle; arginine and proline metabolism
X-18888	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-18889	tryptophan metabolism	-/-	tryptophan metabolism
X-18938	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-19141	androgenic steroids	-/-	androgenic steroids
X-19434	androgenic steroids	-/-	androgenic steroids
X-19561	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-19913	pyrimidine metabolism, uracil containing	-/-	pyrimidine metabolism, uracil containing
X-20624	acetylated peptides	-/-	acetylated peptides
X-21258	tyrosine metabolism	-/-	tyrosine metabolism
X-21283	purine metabolism, adenine containing	-/-	purine metabolism, adenine containing
X-21410	androgenic steroids	-/-	androgenic steroids
X-21467	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-21470	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-21668	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-21737	benzoate metabolism	-/-	benzoate metabolism
X-21785	tyrosine metabolism	-/-	tyrosine metabolism
X-21788	fatty acid synthesis	-/-	methionine, cysteine, sam and taurine metabolism
X-21792	long chain fatty acid	-/-	long chain fatty acid
X-21804	benzoate metabolism	-/-	benzoate metabolism
X-21815	benzoate metabolism	-/-	benzoate metabolism
X-21821	long chain fatty acid	-/-	tryptophan metabolism
X-21827	fatty acid, dicarboxylate, or long chain fatty acid	-/-	long chain fatty acid

Table continued (p. 10/13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-21829	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-21831	long chain fatty acid	-/-	long chain fatty acid
X-21834	tyrosine metabolism	-/-	tyrosine metabolism
X-21842	glutathione metabolism, or methionine, cysteine, sam and taurine metabolism	-/-	glutathione metabolism, or methionine, cysteine, sam and taurine metabolism
X-21846	carnitine metabolism	-/-	—
X-21851	progestin steroids	-/-	progestin steroids
X-22143	pantothenate and coa metabolism	-/-	pantothenate and coa metabolism
X-22158	fatty acid metabolism(acyl carnitine)	-/-	fatty acid metabolism(acyl carnitine)
X-22379	androgenic steroids	-/-	androgenic steroids
X-22475	leucine, isoleucine and valine metabolism	-/-	leucine, isoleucine and valine metabolism
X-22508	chemical	-/-	chemical
X-22757	fatty acid, monohydroxy	-/-	fatty acid, monohydroxy
X-22834	pregnenolone steroids, or progestin steroids	-/-	pregnenolone steroids, or progestin steroids
X-22836	glutamate metabolism	-/-	glutamate metabolism
X-23164	tocopherol metabolism	-/-	tocopherol metabolism
X-23196	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-23291	benzoate metabolism, or food component/plant	-/-	benzoate metabolism, or food component/plant
X-23304	benzoate metabolism	-/-	benzoate metabolism
X-23306	food component/plant	-/-	food component/plant
X-23314	glycolysis, gluconeogenesis, and pyruvate metabolism	-/-	glycolysis, gluconeogenesis, and pyruvate metabolism
X-23423	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-23429	food component/plant	-/-	food component/plant
X-23518	tyrosine metabolism	-/-	tyrosine metabolism
X-23581	glutamate metabolism, or urea cycle; arginine and proline metabolism	-/-	glutamate metabolism, or urea cycle; arginine and proline metabolism
X-23583	glutamate metabolism	-/-	glutamate metabolism
X-23584	glycolysis, gluconeogenesis, and pyruvate metabolism	-/-	glycolysis, gluconeogenesis, and pyruvate metabolism
X-23590	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-23593	gamma-glutamyl amino acid	-/-	gamma-glutamyl amino acid
X-23644	benzoate metabolism	-/-	benzoate metabolism
X-23645	leucine, isoleucine and valine metabolism, or methionine, cysteine, sam and taurine metabolism	-/-	leucine, isoleucine and valine metabolism, or methionine, cysteine, sam and taurine metabolism
X-23647	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-23648	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-23649	chemical, or food component/plant	-/-	chemical, or food component/plant

Table continued (p. 11 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-23653	glutamate metabolism, or leucine, isoleucine and valine metabolism, or polyamine metabolism	-/-	glutamate metabolism, or leucine, isoleucine and valine metabolism, or polyamine metabolism
X-23655	chemical	-/-	chemical
X-23656	polyamine metabolism	-/-	histidine metabolism
X-23659	chemical	-/-	chemical
X-23662	lysine metabolism	-/-	lysine metabolism
X-23668	polyamine metabolism	-/-	histidine metabolism
X-23739	lysine metabolism	-/-	lysine metabolism
X-23747	polyamine metabolism	-/-	polyamine metabolism
X-23776	chemical	-/-	chemical
X-23780	glutamate metabolism	-/-	glutamate metabolism
X-23908	alanine and aspartate metabolism, or glycine, serine and threonine metabolism	-/-	alanine and aspartate metabolism, or glycine, serine and threonine metabolism
X-23997	tyrosine metabolism	-/-	tyrosine metabolism
X-24228	disaccharides and oligosaccharides	-/-	disaccharides and oligosaccharides
X-24249	tryptophan metabolism	-/-	tryptophan metabolism
X-24270	benzoate metabolism	-/-	benzoate metabolism
X-24293	food component/plant	-/-	food component/plant
X-24328	fatty acid synthesis	-/-	fatty acid synthesis
X-24329	tryptophan metabolism	-/-	tryptophan metabolism
X-24330	long chain fatty acid	-/-	long chain fatty acid
X-24338	tryptophan metabolism	-/-	tryptophan metabolism
X-24339	fatty acid metabolism (acyl glutamine)	-/-	fatty acid metabolism (acyl glutamine)
X-24343	food component/plant	-/-	food component/plant
X-24344	benzoate metabolism	-/-	benzoate metabolism
X-24345	acetylated peptides	-/-	acetylated peptides
X-24347	fatty acid metabolism (acyl glutamine)	-/-	—
X-24348	sterol/steroid	-/-	—
X-24349	corticosteroids	-/-	corticosteroids
X-24350	acetylated peptides	-/-	acetylated peptides
X-24353	glutamate metabolism, or leucine, isoleucine and valine metabolism, or polyamine metabolism	-/-	glutamate metabolism, or leucine, isoleucine and valine metabolism, or polyamine metabolism
X-24357	corticosteroids	-/-	corticosteroids
X-24358	long chain fatty acid	-/-	long chain fatty acid
X-24359	tocopherol metabolism	-/-	tocopherol metabolism
X-24360	long chain fatty acid	-/-	long chain fatty acid
X-24361	androgenic steroids	-/-	androgenic steroids
X-24362	progestin steroids	-/-	progestin steroids
X-24387	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism
X-24400	pterin metabolism, or tetrahydrobiopterin metabolism	-/-	pterin metabolism, or tetrahydrobiopterin metabolism

Table continued (p. 12 / 13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-24402	leucine, isoleucine and valine metabolism	-/-	leucine, isoleucine and valine metabolism
X-24408	pyrimidine metabolism, uracil containing	-/-	pyrimidine metabolism, uracil containing
X-24410	glycine, serine and threonine metabolism	-/-	glycine, serine and threonine metabolism
X-24411	glycine, serine and threonine metabolism	-/-	glycine, serine and threonine metabolism
X-24412	benzoate metabolism	-/-	benzoate metabolism
X-24414	chemical	-/-	chemical
X-24416	tocopherol metabolism	-/-	tocopherol metabolism
X-24417	androgenic steroids	-/-	androgenic steroids
X-24422	nicotinate and nicotinamide metabolism	-/-	nicotinate and nicotinamide metabolism
X-24431	ascorbate and aldarate metabolism	-/-	ascorbate and aldarate metabolism
X-24432	ascorbate and aldarate metabolism	-/-	ascorbate and aldarate metabolism
X-24452	histidine metabolism, or lysine metabolism	-/-	histidine metabolism, or lysine metabolism
X-24455	tryptophan metabolism	-/-	tryptophan metabolism
X-24460	tca cycle	-/-	tca cycle
X-24462	pentose metabolism	-/-	pentose metabolism
X-24465	glutamate metabolism	-/-	glutamate metabolism
X-24466	methionine, cysteine, sam and taurine metabolism	-/-	methionine, cysteine, sam and taurine metabolism
X-24468	leucine, isoleucine and valine metabolism	-/-	leucine, isoleucine and valine metabolism
X-24473	food component/plant	-/-	food component/plant
X-24475	food component/plant	-/-	food component/plant
X-24490	phenylalanine metabolism, or tyrosine metabolism	-/-	phenylalanine metabolism, or tyrosine metabolism
X-24494	androgenic steroids, or corticosteroids	-/-	androgenic steroids, or corticosteroids
X-24498	food component/plant	-/-	food component/plant
X-24512	glutathione metabolism	-/-	glutathione metabolism
X-24513	lysine metabolism	-/-	lysine metabolism
X-24514	guanidino and acetamido metabolism	-/-	guanidino and acetamido metabolism
X-24518	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	alanine and aspartate metabolism, or lysine metabolism, or urea cycle; arginine and proline metabolism
X-24519	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-24522	carnitine metabolism, or fatty acid metabolism (also bcaa metabolism), or fatty acid synthesis	-/-	carnitine metabolism, or fatty acid metabolism (also bcaa metabolism), or fatty acid synthesis
X-24542	tyrosine metabolism	-/-	tyrosine metabolism
X-24543	benzoate metabolism	-/-	benzoate metabolism
X-24546	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-24586	alanine and aspartate metabolism	-/-	alanine and aspartate metabolism

Table continued (p. 13/13)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-24587	tyrosine metabolism	-/-	tyrosine metabolism
X-24588	long chain fatty acid	-/-	long chain fatty acid
X-24590	tyrosine metabolism	-/-	tyrosine metabolism
X-24660	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-24669	secondary bile acid metabolism	-/-	secondary bile acid metabolism
X-24686	glycolysis, gluconeogenesis, and pyruvate metabolism	-/-	glycolysis, gluconeogenesis, and pyruvate metabolism
X-24736	lysine metabolism, or urea cycle; arginine and proline metabolism	-/-	lysine metabolism
X-24738	inositol metabolism	-/-	inositol metabolism
X-24757	benzoate metabolism	-/-	benzoate metabolism
X-24758	polyamine metabolism	-/-	polyamine metabolism
X-24760	benzoate metabolism	-/-	benzoate metabolism
X-24766	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-24795	food component/plant	-/-	food component/plant
X-24796	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-24798	carnitine metabolism	-/-	—
X-24799	long chain fatty acid	-/-	long chain fatty acid
X-24807	phospholipid metabolism	-/-	phospholipid metabolism
X-24808	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-24809	urea cycle; arginine and proline metabolism	-/-	urea cycle; arginine and proline metabolism
X-24811	chemical, or food component/plant	-/-	chemical, or food component/plant
X-24834	fibrinogen cleavage peptide	-/-	—
X-24838	fatty acid, dicarboxylate, or ketone bodies	-/-	fatty acid, dicarboxylate, or ketone bodies
X-24860	fatty acid metabolism(acyl carnitine)	-/-	fatty acid metabolism (acyl glutamine)
X-24969	glutathione metabolism, or tryptophan metabolism	-/-	glutathione metabolism, or tryptophan metabolism
X-24971	leucine, isoleucine and valine metabolism	-/-	leucine, isoleucine and valine metabolism
X-24983	chemical	-/-	chemical

Table B.6: Predicted sub pathways of unknown metabolites in SUMMIT/GCKD, SUMMIT, and GCKD

Sub pathways were predicted with three different network models based on: SUMMIT/GCKD, SUMMIT, or GCKD. For 435 of 677 unknown metabolites, sub pathways were predicted in at least one of the three sets. The remaining 242 unknown metabolites (with no sub pathway annotation) are not shown (cf. Table B.5). '-/-' and '—' indicate that the unknown metabolite was not measured in the respective cohort, or that the unknown was measured, but no super pathway was predicted, respectively.

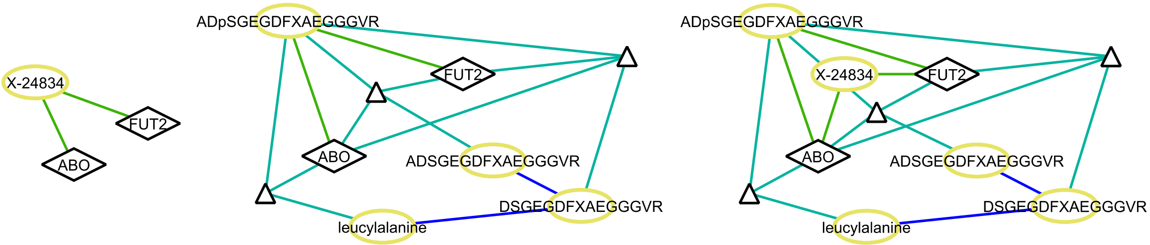
Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-11334	histidine metabolism, or lysine metabolism	lysine metabolism	histidine metabolism
X-11787	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	guanidino and acetamido metabolism
X-12093	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	histidine metabolism, or urea cycle; arginine and proline metabolism
X-12511	urea cycle; arginine and proline metabolism	amino fatty acid, or urea cycle; arginine-, proline-, metabolism	urea cycle; arginine and proline metabolism
X-13835	histidine metabolism	histidine metabolism	histidine metabolism
X-15461	lysine metabolism	phenylalanine & tyrosine metabolism	lysine metabolism
X-14314	glutathione metabolism	—	glutathione metabolism
X-14095	—	—	xanthine metabolism
X-14056	methionine, cysteine, sam and taurine metabolism	hemoglobin and porphyrin metabolism	methionine, cysteine, sam and taurine metabolism
X-01911	tyrosine metabolism	ascorbate and aldarate metabolism	—
X-14302	tryptophan metabolism	polypeptide	—
X-12007	disaccharides and oligosaccharides	fructose, mannose, galactose, starch, and sucrose metabolism	—
X-12206	ascorbate and aldarate metabolism	ascorbate and aldarate metabolism	—
X-17162	hemoglobin and porphyrin metabolism	hemoglobin and porphyrin metabolism	—
X-15492	tca cycle	sterol/steroid	tca cycle
X-11440	androgenic steroids	sterol, steroid	androgenic steroids
X-11470	progestin steroids	sterol/steroid	progestin steroids
X-11540	fatty acid metabolism(acyl carnitine), or phospholipid metabolism	glycerolipid metabolism	fatty acid metabolism(acyl carnitine)
X-13866	fatty acid metabolism (acyl glutamine)	long chain fatty acid	fatty acid metabolism (acyl glutamine)
X-14626	fatty acid, dicarboxylate, or secondary bile acid metabolism	fatty acid, dicarboxylate, or lysolipid	secondary bile acid metabolism
X-14662	secondary bile acid metabolism	bile acid metabolism	secondary bile acid metabolism
X-16654	secondary bile acid metabolism	bile acid metabolism	secondary bile acid metabolism
X-16947	inositol metabolism	inositol metabolism	—
X-17299	carnitine metabolism	carnitine metabolism	—
X-17438	fatty acid, dicarboxylate	fatty acid, dicarboxylate	—
X-12822	fatty acid metabolism (acyl glutamine)	—	fatty acid metabolism (acyl glutamine)
X-12846	androgenic steroids	—	androgenic steroids
X-12860	fatty acid metabolism(acyl carnitine)	—	fatty acid metabolism (acyl glutamine)
X-15486	fatty acid metabolism(acyl glycine)	—	fatty acid metabolism(acyl glycine)
X-02249	essential fatty acid, or fatty acid, dicarboxylate, or long chain fatty acid	essential fatty acid, or fatty acid, dicarboxylate, or long chain fatty acid	tryptophan metabolism
X-16071	fatty acid metabolism (acyl glutamine)	tryptophan metabolism	fatty acid metabolism (acyl glutamine)

Table continued (p. 2/2)

Metabolite	SUMMIT/GCKD	SUMMIT	GCKD
X-15497	fatty acid metabolism (acyl glutamine)	drug	fatty acid metabolism (acyl glutamine)
X-11429	purine metabolism, (hypo)xanthine/inosine containing, or pyrimidine metabolism, uracil containing	purine metabolism, (hypo)xanthine/inosine containing, or pyrimidine metabolism, uracil containing	—
X-12844	nad metabolism	nad metabolism	—
X-12216	acetylated peptides	phenylalanine & tyrosine metabolism	—
X-12730	food component/plant	food component/plant	food component/plant
X-12847	food component/plant	food component/plant	food component/plant
X-13728	xanthine metabolism	xanthine metabolism	xanthine metabolism
X-15728	benzoate metabolism, or food component/plant	benzoate metabolism	food component/plant
X-11452	food component/plant	food component/plant	—
X-12230	benzoate metabolism	benzoate metabolism	—
X-12231	food component/plant	food component/plant	—
X-12329	food component/plant	food component/plant	—
X-12407	food component/plant	food component/plant	—
X-12830	food component/plant	food component/plant	—
X-17185	benzoate metabolism	benzoate metabolism	—
X-09789	chemical	—	chemical
X-12212	food component/plant	—	food component/plant
X-12543	benzoate metabolism	phenylalanine & tyrosine metabolism	benzoate metabolism
X-17685	benzoate metabolism	phenylalanine & tyrosine metabolism	benzoate metabolism
X-12704	benzoate metabolism	phenylalanine & tyrosine metabolism	—

Table B.7: Predicted sub pathways of unknown metabolites measured in SUMMIT and GCKD

Sub pathways were predicted in 51 of 59 unknown metabolites that were measured in SUMMIT and GCKD. Entries are ordered according to respective super pathway predictions of Table 4.5



(a) GCKD (b) SUMMIT (c) GCKD/SUMMIT

Figure B.4: Neighbors of X-24834 in sets of GCKD and SUMMIT
(a) X-24834 has been measured in the GCKD cohort and is only connected to two genes. (b) In the network model of SUMMIT these two genes are connected to several di- or poly-peptides. (c) In consequence, X-24834 is neighbor of these metabolites in the network model of the merged set SUMMIT/GCKD.

Manual selection of candidate molecules

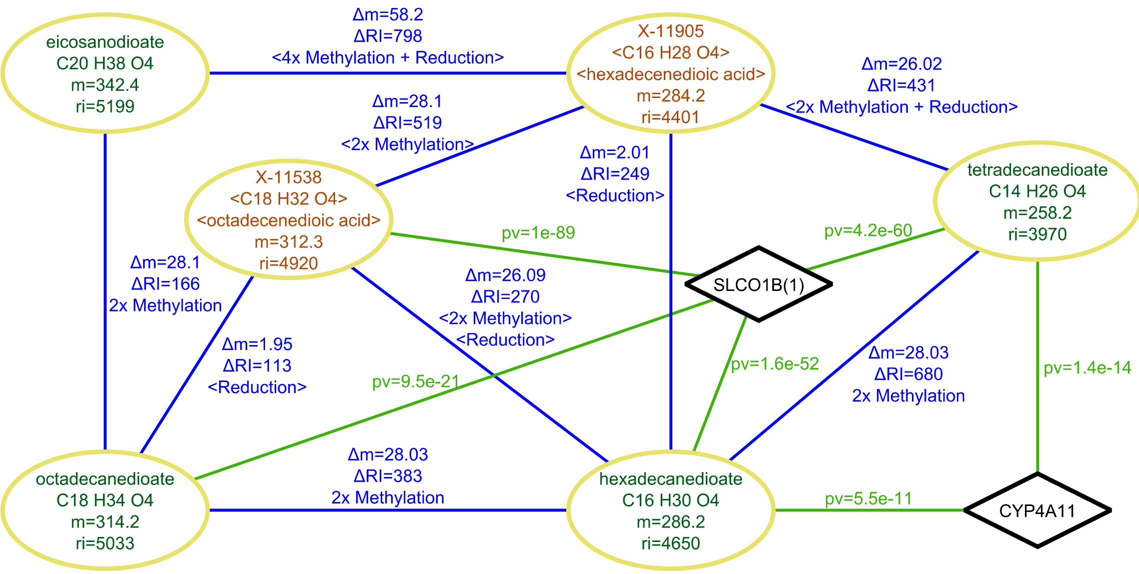


Figure B.5: Neighboring metabolites of X-11905 and X-11538
Predicted super and sub pathways and predicted reactions with neighboring metabolites lead to the candidate molecule hexadecenedioic acid and octadecenedioic acid for X-11905 and X-11583, respectively. Combined reactions, e.g. methylation + reduction, were added manually. Information in '<>' was predicted.

Metabolite	R	Candidate or isomeric compound	Evidence
X-11334	H	(1-Ribosylimidazole)-4-acetate	neighbor of Methylimidazoleacetic acid and pipecolate; no direct reaction, but close in pathway
X-11540	H	trans-2-Dodecenoylcarnitine	similar in maas, structure and RI of neighboring metabolite octanoylcarnitine
X-11787	H	hydroxyleucine or hydrosyisoleucine	close to amino acid PW
X-12212	H	isobar of X-21834 isobar of 4-vinylguaiacol sulfate	X-12212 (m/z=229.0179) and X-21834 (m/z=299.0184); similar m/z and small delta RI
		[4-(3-oxopropyl)phenyl]oxidanesulfonic acid [2-hydroxy-5-(prop-2-en-1-yl)phenyl]oxidanesulfonic acid 4-[(1E)-3-hydroxyprop-1-en-1-yl]phenyloxidanesulfonic acid [4-(3-hydroxyprop-1-en-1-yl)phenyl]oxidanesulfonic acid	candidates have similar structure to neighbors; unknowns are not 4-vinylguaiacol sulfate, because it is measured and isobaric
X-12230	H	Tyrosol 4-sulfate (5-ethyl-2-hydroxyphenyl)oxidanesulfonic acid	neighbor of 4-ethylpenylsulfate, second-level-neighbor of 4-vinylphenolsulfate
X-12511	H	Valproylglycine Capryloylglycine	similar mass, structure of neighboring measured metabolites (especially N-acetylleucine, N-acetylasparagine, N-alpha-acetylornithine, N-acetylarginine, N-acetylcitrulline, of for the fatty acid: 2-hydroxyoctanoate)
X-12543	H	isobar of 3-(4-hydroxyphenyl)lactate Dihydroxyphenyl-propanoic acid Hydroxyphenyllactate Dihydroxyhydrocinnamic acid hydroxy-(methoxyphenyl)acetic acid Dimethoxybenzoic acid Homovanillic acid	direct neighbors; same mass; very small delta RI of 30; cluster member of structurally similar molecules
X-12734	E	isobar of 2-methoxyresorcinol sulfate isobar of 3-methoxycatechol sulfate (1)	same maas and similar RI; neighbors over another unknown metabolite
X-12749	E	O-Phosphotyrosine	neighbor of O-sulfo-L-tyrosine similar mass, but measured on different platform
X-12847	H		Hint: neighbor of thymol sulfate, and very close m/z and RI
X-12860	H	Hydroxyhexanoycarnitine	neighbor of structurally similar 3-hydroxybutyrylcarnitine (2); similar mass and small delta RI
X-13431	H	isobar of Nonanoylcarnitine 2,6 Dimethylheptanoyl carnitine	neighbor of Nonanoylcarnitine through ACADL; same mass; unknown is not Nonanoylcarnitine, because metabolite is annotated
X-13728	H	6-amino-5[N-methylformylamino]-1-methyluracil	surroundet by similar molecules, equal mass
X-13834	H	nonenedioic acid/ nonenedioate	neighbor of 4-octenedioate, delta RI=182
X-15461	H	5-Acetamidovalerate (N-acetyl-cadaverine)	similar structure of neighbors 2-piperidinone, 5-aminovalerate, N-acetyl-cadaverine; small delta RI, same mass
X-15469	H	isobar of heptanoylglutamine	neighbors over another unknown; same mass: delta m=0.0008 and small delta RI=145

Table continued (p. 2/2)

Metabolite	R	Candidate or isomeric compound	Evidence
X-15636	E	isobar of eugenol sulfate	same mass; small delta RI
X-15728	H	isobar of X-24343 (4-ethyl-2-methoxyphenyl)oxidanesulfonic acid	X-15728 (m/z=231.0334) and X-24343 (m/z=231.0331); similar m/z and very small delta RI
X-16071	H	1H-Indole-3-carboxaldehyde	in neighborhood of structurally similar metabolites; similar mass and RI
X-16124	H	O-methoxycatechol-O-sulphate	neighbor over another unknown molecule to eugenol sulfate
X-16940	E	Pyrogallol-1-O-sulphate Pyrogallol-2-O-sulphate	neighbor over another unknown to structurally similar molecules
X-17162	H		Hint: neighbor of L-urobilin; similar m/z and RI
X-17185	H	(5-ethenyl-2-hydroxyphenyl)oxidanesulfonic acid	neighbor of 4-vinylphenolsulfate; equal mass; very small delta RI
X-17337	H	(9E)-10-nitrooctadecenoic acid (9E)-9-nitrooctadecenoic acid	similar mass, structure and RI of neighboring (over another unknown) metabolite octanoylcarnitine
X-17342	H	2-[4-hydroxy-3-(sulfooxy)phenyl]acetic acid	neighbor of structurally similar 4-hydroxymandelate
X-17685	H	(4-ethyl-2,6-dihydroxyphenyl)oxidanesulfonic acid	sourroundet by structurally similar molecules
X-21792	H	Methylene suberic acid	neighbor of structurally similar stearate, palmitate, arachidonate, and adrenate through a common associated gene, ACOT2
X-21834	H	*cf. isobaric X-12212*	
X-23423	H	similar to 2-Aminoheptanedioic acid	neighbor of N-delta-acetylornithine; delta mass of 1 and delta RI of 50
X-23641	H	isobar of L-Octanoylcarnitine	connected over 4 other nodes to L-Octanoylcarnitine; equal m/z; delta RI=40; similar metabolites in direct neighborhood
X-24343	H	*cf. isobaric X-15728*	
X-24357	H	urocortisone	small delta RI and similar structure to neighboring cortisol
X-24983	H	Ne,Ne dimethyllysine	similar mass, structure and RI of neighboring (over another unknown) metabolites (especially N-acetylornithine, 2-aminooctanoic acid, homocitrulline, N-methylpipecolate, N6-carboxyethyllysine); dimethyl-lysine verified by Metabolon$^{\text{TM}}$

Table B.8: Manually selected candidate molecules basted on HMDB compounds 36 specific candidate molecules were selected for 33 unknown metabolites based on predicted super and sub pathways, predicted reactions, database searches of the absolute masses and structural similarity to neighboring (according to the network model) metabolites. The second column resolution (R) indicates if unknown metabolites were measured on the established (E) or on the new high-resolution (H) platform of Metabolon$^{\text{TM}}$.

Automated selection of candidate molecules

V.	Run	Fingerp. Tanimoto	MCS Tanimoto	MCS Overlap	Sensitivity	Specificity	Mol.	True	False
1	1	0.686	0.647	0.908	0.866	0.607	48	47	84
1	2	0.686	0.659	0.904	0.868	0.617	45	41	76
1	3	0.742	0.670	0.905	0.851	0.656	43	40	68
1	4	0.749	0.677	0.862	0.876	0.636	34	33	68
1	5	0.693	0.681	0.910	0.866	0.639	40	33	92
1	6	0.736	0.656	0.900	0.850	0.650	42	40	77
1	7	0.680	0.645	0.904	0.859	0.612	44	43	85
1	8	0.686	0.657	0.908	0.869	0.620	47	44	101
1	9	0.686	0.668	0.904	0.868	0.609	38	34	43
1	10	0.686	0.673	0.907	0.865	0.630	37	34	70
2	1	0.586	0.647	0.908	0.896	0.533	49	48	89
2	2	0.586	0.659	0.904	0.900	0.544	46	41	84
2	3	0.742	0.670	0.805	0.881	0.587	46	43	112
2	4	0.649	0.677	0.862	0.893	0.590	34	34	78
2	5	0.543	0.681	0.910	0.912	0.524	40	36	126
2	6	0.586	0.656	0.900	0.894	0.542	45	44	97
2	7	0.580	0.645	0.904	0.893	0.541	44	43	97
2	8	0.586	0.657	0.908	0.899	0.552	49	45	121
2	9	0.586	0.668	0.904	0.897	0.538	40	35	51
2	10	0.586	0.673	0.907	0.899	0.562	38	35	89

Table B.9: Parameters determined during cross-validation of candidate selection During each run of the 10-fold cross-validation cutoffs for parameters fingerprint Tanimoto, MCS Tanimoto and MCS Overlap were determined based on the maximum of sensitivity and specificity, while sensitivity was weighted twice for variant 1 and three-fold for variant 2, while specificity was weighted once. The last three columns describe the number of test-molecules and the number of true or false selected candidates.

Unknown	Candidate	PubChem CID	△Mass	MFTC	MSTC	MSOC
X-02249	Benzo[a]pyrene-4,5-oxide	37786	0.0351	0.64*	0.48	0.85
	Benzo[a]pyrene-9,10-oxide	37456	0.0351	0.64*	0.47	0.77
X-11261	octenoyl carnitine	53481667	0.0068	0.61*	0.5	0.91*
X-11478	Perillic acid	1256	0.0073	0.61*	0.53	0.75
	(4-hydroxy-3-methoxyphenyl)acetaldehyde	151276	0.0291	0.6*	0.53	0.73
X-11787	L-4-hydroxyglutamic semialdehyde	25201126	0.0436	0.62*	0.64	0.9*
X-12093	L-alanyl-L-leucine	6992388	0.0071	0.94*	0.86*	1*
X-12100	5-hydroxy-L-tryptophan	439280	0.0071	0.67*	0.33	0.58
X-12170	trans-caffeate	689043	0.0185	0.58*	0.65	0.85
	3-Hydroxykynurenamine	440736	0.0291	0.84*	0.65	0.85
	5-Hydroxykynurenamine	164719	0.0291	0.79*	0.56	0.77
	adrenochrome o-semiquinone	10313383	0.0053	0.61*	0.5	0.69
X-12543	3-Methoxy-4-hydroxyphenylglycolaldehyde	440729	0.0069	0.77*	0.62	0.77
X-12688	L-alanyl-L-leucine	6992388	0.0073	0.69*	0.5	0.8

Evidence
neighbor of 16alpha-Hydroxyestrone (delta mass=18.0681, Structure: Tanimoto Coefficient=0.355, Overlap Coefficient=0.524, Fingerprint: Tanimoto Coefficient=0.6); neighbor of 2-Hydroxyestradiol-17beta (delta mass=20.0837, Structure: Tanimoto Coefficient=0.312, Overlap Coefficient=0.476, Fingerprint: Tanimoto Coefficient=0.61); neighbor of estradiol (delta mass=4.0888, Structure: Tanimoto Coefficient=0.367, Overlap Coefficient=0.55, Fingerprint: Tanimoto Coefficient=0.644); neighbor of estrone (delta mass=2.0732, Structure: Tanimoto Coefficient=0.367, Overlap Coefficient=0.55, Fingerprint: Tanimoto Coefficient=0.6)
neighbor of 16alpha-Hydroxyestrone (delta mass=18.0681, Structure: Tanimoto Coefficient=0.355, Overlap Coefficient=0.524, Fingerprint: Tanimoto Coefficient=0.599); neighbor of 2-Hydroxyestradiol-17beta (delta mass=20.0837, Structure: Tanimoto Coefficient=0.312, Overlap Coefficient=0.476, Fingerprint: Tanimoto Coefficient=0.609); neighbor of estradiol (delta mass=4.0888, Structure: Tanimoto Coefficient=0.367, Overlap Coefficient=0.55, Fingerprint: Tanimoto Coefficient=0.644); neighbor of estrone (delta mass=2.0732, Structure: Tanimoto Coefficient=0.367, Overlap Coefficient=0.55, Fingerprint: Tanimoto Coefficient=0.599)
neighbor of carnitine (delta mass=124.0888, Structure: Tanimoto Coefficient=0.476, Overlap Coefficient=0.909, Fingerprint: Tanimoto Coefficient=0.474); neighbor of suberoylcarnitine (C8-DC) (delta mass=31.9971, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.7, Fingerprint: Tanimoto Coefficient=0.605)
neighbor of 3-phenylpropionate (hydrocinnamate) (delta mass=15.9994, Structure: Tanimoto Coefficient=0.438, Overlap Coefficient=0.636, Fingerprint: Tanimoto Coefficient=0.607)
neighbor of 3-phenylpropionate (hydrocinnamate) (delta mass=15.963, Structure: Tanimoto Coefficient=0.533, Overlap Coefficient=0.727, Fingerprint: Tanimoto Coefficient=0.6)
neighbor of N-acetylglutamine (delta mass=41.0192, Structure: Tanimoto Coefficient=0.643, Overlap Coefficient=0.9, Fingerprint: Tanimoto Coefficient=0.616)
neighbor of homocitrulline (delta mass=13.0131, Structure: Tanimoto Coefficient=0.421, Overlap Coefficient=0.615, Fingerprint: Tanimoto Coefficient=0.71); neighbor of N-acetylarginine (delta mass=13.9978, Structure: Tanimoto Coefficient=0.611, Overlap Coefficient=0.786, Fingerprint: Tanimoto Coefficient=0.736); neighbor of N-acetylasparagine (delta mass=28.0677, Structure: Tanimoto Coefficient=0.625, Overlap Coefficient=0.833, Fingerprint: Tanimoto Coefficient=0.739); neighbor of N-acetylcitrulline (delta mass=14.9818, Structure: Tanimoto Coefficient=0.611, Overlap Coefficient=0.786, Fingerprint: Tanimoto Coefficient=0.786); neighbor of N-acetylglutamine (delta mass=14.0593, Structure: Tanimoto Coefficient=0.688, Overlap Coefficient=0.846, Fingerprint: Tanimoto Coefficient=0.783); neighbor of N-acetylleucine (delta mass=29.0338, Structure: Tanimoto Coefficient=0.857, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.935); neighbor of N-acetylmethionine sulfoxide (delta mass=4.9321, Structure: Tanimoto Coefficient=0.588, Overlap Coefficient=0.769, Fingerprint: Tanimoto Coefficient=0.662); neighbor of N-alpha-acetylornithine (delta mass=28.0313, Structure: Tanimoto Coefficient=0.733, Overlap Coefficient=0.917, Fingerprint: Tanimoto Coefficient=0.831); neighbor of N-delta-acetylornithine (delta mass=28.0385, Structure: Tanimoto Coefficient=0.444, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.788); neighbor of N-methylpipecolate (delta mass=59.0298, Structure: Tanimoto Coefficient=0.6, Overlap Coefficient=0.9, Fingerprint: Tanimoto Coefficient=0.629); neighbor of N2-acetyllysine (delta mass=14.0229, Structure: Tanimoto Coefficient=0.688, Overlap Coefficient=0.846, Fingerprint: Tanimoto Coefficient=0.848)
neighbor of quinolinate (delta mass=53.0629, Structure: Tanimoto Coefficient=0.333, Overlap Coefficient=0.583, Fingerprint: Tanimoto Coefficient=0.665)
neighbor of xanthurenate (delta mass=24.9952, Structure: Tanimoto Coefficient=0.647, Overlap Coefficient=0.846, Fingerprint: Tanimoto Coefficient=0.584)
neighbor of kynurenate (delta mass=8.9527, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.692, Fingerprint: Tanimoto Coefficient=0.682); neighbor of xanthurenate (delta mass=24.9476, Structure: Tanimoto Coefficient=0.647, Overlap Coefficient=0.846, Fingerprint: Tanimoto Coefficient=0.841)
neighbor of kynurenate (delta mass=8.9527, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.692, Fingerprint: Tanimoto Coefficient=0.693); neighbor of xanthurenate (delta mass=24.9476, Structure: Tanimoto Coefficient=0.556, Overlap Coefficient=0.769, Fingerprint: Tanimoto Coefficient=0.794)
neighbor of xanthurenate (delta mass=24.9714, Structure: Tanimoto Coefficient=0.474, Overlap Coefficient=0.692, Fingerprint: Tanimoto Coefficient=0.607)
neighbor of 3-(3-hydroxyphenyl)propionate (delta mass=16.0022, Structure: Tanimoto Coefficient=0.562, Overlap Coefficient=0.75, Fingerprint: Tanimoto Coefficient=0.711); neighbor of 3-(4-hydroxyphenyl)lactate (delta mass=0, Structure: Tanimoto Coefficient=0.625, Overlap Coefficient=0.769, Fingerprint: Tanimoto Coefficient=0.773)
neighbor of 4-acetamidobutanoate (delta mass=57.0578, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.688); neighbor of acisoga (delta mass=18.0178, Structure: Tanimoto Coefficient=0.421, Overlap Coefficient=0.615, Fingerprint: Tanimoto Coefficient=0.588)

Table continued (p. 2/5)

Unknown	Candidate	PubChem CID	△Mass	MFTC	MSTC	MSOC
X-12860	glutaryl carnitine	53481622	0.0432	0.87*	0.85*	1*
X-13684	3-mercaptolactate-cysteine disulfide	193536	0.0068	0.73*	0.5	1*
X-13866	N-Ribosylnicotinamide	439924	0.018	0.66*	0.3	0.58
X-15497	fructoseglycine	3081391	0.0081	0.61*	0.5	0.69
	salsoline-1-carboxylate	320322	0.0071	0.61*	0.58	0.85
X-15503	dopachrome o-semiquinone	53481550	0.0309	0.62*	0.65	0.79
X-16071	6-amino-2-oxohexanoic acid	439954	0.0284	0.59*	0.56	0.9*
	L-allysine	207	0.0284	0.63*	0.47	0.8
	2-oxoglutaramate	48	0.008	0.6*	0.53	0.9*
X-16563	2-acetamido-5-oxopentanoate	192878	0.0295	0.83*	0.71*	1*
X-17323	16-hydroxypalmitate	7058075	0.0422	0.64*	0.32	0.58
X-18888	2-keto-3-deoxy-D-glycero-D-galactononic acid	22833524	0.0446	0.79*	0.57	0.8
X-21831	12-oxo-20-carboxy-leukotriene B4	53481457	0.021	0.7*	0.32	0.82

Evidence
neighbor of 3-hydroxybutyrylcarnitine (2) (delta mass=27.9876, Structure: Tanimoto Coefficient=0.8, Overlap Coefficient=0.941, Fingerprint: Tanimoto Coefficient=0.867); neighbor of acetylcarnitine (C2) (delta mass=72.0211, Structure: Tanimoto Coefficient=0.737, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.831); neighbor of decanoylcarnitine (C10) (delta mass=40.1041, Structure: Tanimoto Coefficient=0.708, Overlap Coefficient=0.895, Fingerprint: Tanimoto Coefficient=0.765); neighbor of hexanoylcarnitine (delta mass=15.9513, Structure: Tanimoto Coefficient=0.85, Overlap Coefficient=0.944, Fingerprint: Tanimoto Coefficient=0.839); neighbor of octanoylcarnitine (C8) (delta mass=12.0728, Structure: Tanimoto Coefficient=0.773, Overlap Coefficient=0.895, Fingerprint: Tanimoto Coefficient=0.788)
neighbor of aspartate (delta mass=107.9704, Structure: Tanimoto Coefficient=0.353, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.603); neighbor of cysteine (delta mass=119.9881, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.731); neighbor of erythro-4-hydroxy-L-glutamate(1-) (delta mass=77.9598, Structure: Tanimoto Coefficient=0.316, Overlap Coefficient=0.545, Fingerprint: Tanimoto Coefficient=0.692)
neighbor of indolelactate (delta mass=49.0237, Structure: Tanimoto Coefficient=0.269, Overlap Coefficient=0.467, Fingerprint: Tanimoto Coefficient=0.648); neighbor of Pyridoxal (delta mass=87.032, Structure: Tanimoto Coefficient=0.304, Overlap Coefficient=0.583, Fingerprint: Tanimoto Coefficient=0.656)
neighbor of heptanoylglutamine (delta mass=21.0658, Structure: Tanimoto Coefficient=0.478, Overlap Coefficient=0.688, Fingerprint: Tanimoto Coefficient=0.593); neighbor of hexanoylglutamine (delta mass=7.0501, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.688, Fingerprint: Tanimoto Coefficient=0.607)
neighbor of ferulic acid 4-sulfate (delta mass=36.9073, Structure: Tanimoto Coefficient=0.4, Overlap Coefficient=0.588, Fingerprint: Tanimoto Coefficient=0.609)
neighbor of kynurenate (delta mass=5.0027, Structure: Tanimoto Coefficient=0.647, Overlap Coefficient=0.786, Fingerprint: Tanimoto Coefficient=0.621)
neighbor of hexanoylglutamine (delta mass=99.0611, Structure: Tanimoto Coefficient=0.35, Overlap Coefficient=0.7, Fingerprint: Tanimoto Coefficient=0.586)
neighbor of heptanoylglutamine (delta mass=113.0768, Structure: Tanimoto Coefficient=0.4, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.616); neighbor of hexanoylglutamine (delta mass=99.0611, Structure: Tanimoto Coefficient=0.421, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.634)
neighbor of heptanoylglutamine (delta mass=113.1132, Structure: Tanimoto Coefficient=0.474, Overlap Coefficient=0.9, Fingerprint: Tanimoto Coefficient=0.583); neighbor of hexanoylglutamine (delta mass=99.0975, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.9, Fingerprint: Tanimoto Coefficient=0.6)
neighbor of 2-methylbutyrylglycine (delta mass=13.9793, Structure: Tanimoto Coefficient=0.533, Overlap Coefficient=0.727, Fingerprint: Tanimoto Coefficient=0.719); neighbor of 3-methylcrotonylglycine (delta mass=15.9876, Structure: Tanimoto Coefficient=0.533, Overlap Coefficient=0.727, Fingerprint: Tanimoto Coefficient=0.583); neighbor of Ammonium (delta mass=156.0423, Structure: Tanimoto Coefficient=0.083, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.066); neighbor of glutaroyl carnitine (delta mass=102.0754, Structure: Tanimoto Coefficient=0.348, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.6); neighbor of glycine (delta mass=98.0368, Structure: Tanimoto Coefficient=0.417, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.443); neighbor of heptanoylglutamine (delta mass=85.0819, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.803); neighbor of hexanoylglutamine (delta mass=71.0662, Structure: Tanimoto Coefficient=0.706, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.826); neighbor of hexanoylglycine (delta mass=0.0291, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.738); neighbor of isobutyrylglycine (delta mass=27.9949, Structure: Tanimoto Coefficient=0.571, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.698); neighbor of isovalerylglycine (delta mass=13.9793, Structure: Tanimoto Coefficient=0.533, Overlap Coefficient=0.727, Fingerprint: Tanimoto Coefficient=0.719); neighbor of N-acetylglycine (delta mass=56.0189, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.656); neighbor of pyroglutamine (delta mass=45.0029, Structure: Tanimoto Coefficient=0.615, Overlap Coefficient=0.889, Fingerprint: Tanimoto Coefficient=0.586); neighbor of tigloylglycine (delta mass=15.9876, Structure: Tanimoto Coefficient=0.533, Overlap Coefficient=0.727, Fingerprint: Tanimoto Coefficient=0.589)
neighbor of suberoylcarnitine (C8-DC) (delta mass=45.9638, Structure: Tanimoto Coefficient=0.323, Overlap Coefficient=0.526, Fingerprint: Tanimoto Coefficient=0.644)
neighbor of 3-hydroxysebacate (delta mass=49.9713, Structure: Tanimoto Coefficient=0.571, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.789)
neighbor of 11,12-EET (delta mass=43.9534, Structure: Tanimoto Coefficient=0.256, Overlap Coefficient=0.435, Fingerprint: Tanimoto Coefficient=0.662); neighbor of 14,15-EET (delta mass=43.9534, Structure: Tanimoto Coefficient=0.225, Overlap Coefficient=0.391, Fingerprint: Tanimoto Coefficient=0.649); neighbor of 5,6-EET (delta mass=43.9534, Structure: Tanimoto Coefficient=0.225, Overlap Coefficient=0.391, Fingerprint: Tanimoto Coefficient=0.697); neighbor of arachidonate (20:4n6) (delta mass=59.9484, Structure: Tanimoto Coefficient=0.231, Overlap Coefficient=0.409, Fingerprint: Tanimoto Coefficient=0.652)

Table continued (p. 3/5)

Unknown	Candidate	PubChem CID	△Mass	MFTC	MSTC	MSOC
X-22158	hexanoyl carnitine	6426853	0.0068	1*	1*	1*
X-22379	5alpha-Dihydrotestosterone glucuronide	44263365	0.0063	0.96*	0.78*	0.88
	androsterone 3-glucosiduronic acid	114833	0.0063	1*	1*	1*
X-22836	6-amino-2-oxohexanoic acid	439954	0.0073	0.59*	0.5	0.7
	L-allysine	207	0.0073	0.64*	0.62	0.8
X-23196	5-L-Glutamyl-L-alanine	440103	0.0177	0.94*	0.6	0.8
	gama-L-glutamyl-L-alanine	11183554	0.0177	0.94*	0.6	0.8
	L-leucyl-L-serine	40489070	0.0187	0.73*	0.6	0.8
	N-acetylserotonin	903	0.0025	0.79*	0.48	0.69
X-23423	2-Amino-3-oxoadipate	440714	0.0289	0.65*	0.56	0.75
X-23581	1-piperideine-6-carboxylate	45266761	0.0073	0.57	0.67*	0.89
X-23583	D-proline	8988	0.0074	0.61*	0.6	0.75
	acetamidopropanal	5460495	0.0074	0.62*	0.45	0.62
X-23593	N-(omega)-Hydroxyarginine	440849	0.0043	0.61*	0.43	0.69
X-23647	Glycylproline	79101	0.0294	0.79*	0.75*	1*
	L-Prolinylglycine	6426709	0.0294	0.74*	0.75*	1*
X-23747	N1,N8-diacetylspermidine	389613	0.0074	0.98*	0.71*	1*
X-23776	6-amino-2-oxohexanoic acid	439954	0.0435	0.61*	0.67*	0.8
	L-allysine	207	0.0435	0.66*	0.67*	0.8
X-24330	Porphobilinogen	1021	0.0187	0.42	0.42	1*

Evidence
neighbor of hexanoylcarnitine (delta mass=0.0072, Structure: Tanimoto Coefficient=1, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=1)
neighbor of epiandrosterone glucuronide (delta mass=0.0073, Structure: Tanimoto Coefficient=0.784, Overlap Coefficient=0.879, Fingerprint: Tanimoto Coefficient=0.961); neighbor of etiocholanolone glucuronide (delta mass=0, Structure: Tanimoto Coefficient=0.784, Overlap Coefficient=0.879, Fingerprint: Tanimoto Coefficient=0.961)
neighbor of epiandrosterone glucuronide (delta mass=0.0073, Structure: Tanimoto Coefficient=1, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=1); neighbor of etiocholanolone glucuronide (delta mass=0, Structure: Tanimoto Coefficient=1, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=1)
neighbor of N-methylglutamate (delta mass=15.9876, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.7, Fingerprint: Tanimoto Coefficient=0.587)
neighbor of N-methylglutamate (delta mass=15.9876, Structure: Tanimoto Coefficient=0.615, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.641)
neighbor of heptanoylglutamine (delta mass=40.0604, Structure: Tanimoto Coefficient=0.571, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.912); neighbor of hexanoylglutamine (delta mass=26.0447, Structure: Tanimoto Coefficient=0.6, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.939)
neighbor of heptanoylglutamine (delta mass=40.0604, Structure: Tanimoto Coefficient=0.571, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.912); neighbor of hexanoylglutamine (delta mass=26.0447, Structure: Tanimoto Coefficient=0.6, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.939)
neighbor of heptanoylglutamine (delta mass=40.024, Structure: Tanimoto Coefficient=0.571, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.709); neighbor of hexanoylglutamine (delta mass=26.0083, Structure: Tanimoto Coefficient=0.6, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.727)
neighbor of indolelactate (delta mass=13.0389, Structure: Tanimoto Coefficient=0.476, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.789); neighbor of Pyridoxal (delta mass=51.0473, Structure: Tanimoto Coefficient=0.333, Overlap Coefficient=0.583, Fingerprint: Tanimoto Coefficient=0.644)
neighbor of homocitrulline (delta mass=14.0705, Structure: Tanimoto Coefficient=0.562, Overlap Coefficient=0.75, Fingerprint: Tanimoto Coefficient=0.647); neighbor of N-delta-acetylornithine (delta mass=0.9549, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.623)
neighbor of N-methylglutamate (delta mass=33.9982, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=0.889, Fingerprint: Tanimoto Coefficient=0.53)
neighbor of N-methyl-GABA (delta mass=2.023, Structure: Tanimoto Coefficient=0.6, Overlap Coefficient=0.75, Fingerprint: Tanimoto Coefficient=0.614)
neighbor of N-methyl-GABA (delta mass=2.023, Structure: Tanimoto Coefficient=0.455, Overlap Coefficient=0.625, Fingerprint: Tanimoto Coefficient=0.625)
neighbor of gamma-glutamylisoleucine (delta mass=70.0379, Structure: Tanimoto Coefficient=0.409, Overlap Coefficient=0.692, Fingerprint: Tanimoto Coefficient=0.605); neighbor of gamma-glutamylvaline (delta mass=56.0223, Structure: Tanimoto Coefficient=0.429, Overlap Coefficient=0.692, Fingerprint: Tanimoto Coefficient=0.613)
neighbor of pro-hydroxy-pro (delta mass=56.0335, Structure: Tanimoto Coefficient=0.75, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.792)
neighbor of pro-hydroxy-pro (delta mass=56.0335, Structure: Tanimoto Coefficient=0.75, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.74)
neighbor of acisoga (delta mass=45.0651, Structure: Tanimoto Coefficient=0.611, Overlap Coefficient=0.846, Fingerprint: Tanimoto Coefficient=0.848); neighbor of N-acetylputrescine (delta mass=99.0684, Structure: Tanimoto Coefficient=0.562, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.897); neighbor of N1,N12-diacetylspermine (delta mass=57.0578, Structure: Tanimoto Coefficient=0.714, Overlap Coefficient=0.938, Fingerprint: Tanimoto Coefficient=0.975)
neighbor of N-methylpipecolate (delta mass=1.972, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.607)
neighbor of N-methylpipecolate (delta mass=1.972, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.661)
neighbor of propionate (delta mass=152.0586, Structure: Tanimoto Coefficient=0.312, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.137)

Table continued (p. 4/5)

Unknown	Candidate	PubChem CID	△Mass	MFTC	MSTC	MSOC
X-24339	Xanthosine 5'-phosphate	73323	0.0336	0.94*	0.73*	0.92*
X-24361	16-Glucuronide-estriol	122281	0.0291	0.63*	0.65	0.79
	testosterone 3-glucosiduronic acid	108192	0.0073	0.84*	0.74*	0.85
X-24410	D-proline	8988	0.0072	0.62*	0.7*	0.88
	acetamidopropanal	5460495	0.0072	0.67*	0.67*	0.86
X-24411	D-proline	8988	0.0072	0.62*	0.64	0.88
	acetamidopropanal	5460495	0.0072	0.63*	0.5	0.75
X-24417	16-Glucuronide-estriol	122281	0.0287	0.63*	0.65	0.79
	testosterone 3-glucosiduronic acid	108192	0.0077	0.84*	0.74*	0.85
X-24452	3-Hydroxy-N6,N6,N6-trimethyl-L-lysine	439460	0.0071	0.81*	0.93*	1*
X-24455	N-acetyl-5-methoxykynuramine	390658	0.0294	0.71*	0.52	0.8
	N-formyl-L-kynurenine	910	0.007	1*	1*	1*
X-24498	cis-beta-D-Glucosyl-2-hydroxycinnamate	5316113	0.0078	0.64*	0.27	0.53
X-24514	putreanine	53477800	0.044	0.59*	0.35	0.55
X-24736	Glycylleucine	92843	0.0184	0.71*	0.5	0.73
	Leucylglycine	97364	0.0184	0.68*	0.41	0.64

Evidence
neighbor of 2,4-decadienoylcoa (delta mass=553.1777, Structure: Tanimoto Coefficient=0.258, Overlap Coefficient=0.708, Fingerprint: Tanimoto Coefficient=0.698); neighbor of 3-decenoylcoa (delta mass=555.1933, Structure: Tanimoto Coefficient=0.258, Overlap Coefficient=0.708, Fingerprint: Tanimoto Coefficient=0.7); neighbor of dADP (delta mass=46.9925, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=0.833, Fingerprint: Tanimoto Coefficient=0.791); neighbor of dATP (delta mass=126.9588, Structure: Tanimoto Coefficient=0.588, Overlap Coefficient=0.833, Fingerprint: Tanimoto Coefficient=0.787); neighbor of dGDP (delta mass=62.9874, Structure: Tanimoto Coefficient=0.7, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.92); neighbor of dGTP (delta mass=142.9537, Structure: Tanimoto Coefficient=0.618, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.914); neighbor of GDP (delta mass=78.9823, Structure: Tanimoto Coefficient=0.733, Overlap Coefficient=0.917, Fingerprint: Tanimoto Coefficient=0.94); neighbor of GTP (delta mass=158.9486, Structure: Tanimoto Coefficient=0.647, Overlap Coefficient=0.917, Fingerprint: Tanimoto Coefficient=0.934)
neighbor of etiocholanolone glucuronide (delta mass=2.052, Structure: Tanimoto Coefficient=0.65, Overlap Coefficient=0.788, Fingerprint: Tanimoto Coefficient=0.632) neighbor of etiocholanolone glucuronide (delta mass=2.0157, Structure: Tanimoto Coefficient=0.737, Overlap Coefficient=0.848, Fingerprint: Tanimoto Coefficient=0.836)
neighbor of 2-piperidinone (delta mass=15.9876, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=0.857, Fingerprint: Tanimoto Coefficient=0.574); neighbor of glutarate (pentanedioate) (delta mass=16.9717, Structure: Tanimoto Coefficient=0.7, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.414); neighbor of N-acetylserine (delta mass=31.9971, Structure: Tanimoto Coefficient=0.636, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.621) neighbor of 2-piperidinone (delta mass=15.9876, Structure: Tanimoto Coefficient=0.667, Overlap Coefficient=0.857, Fingerprint: Tanimoto Coefficient=0.673); neighbor of N-acetylserine (delta mass=31.9971, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.75, Fingerprint: Tanimoto Coefficient=0.632)
neighbor of N-acetylserine (delta mass=31.9971, Structure: Tanimoto Coefficient=0.636, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.621) neighbor of N-acetylserine (delta mass=31.9971, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.75, Fingerprint: Tanimoto Coefficient=0.632)
neighbor of epiandrosterone glucuronide (delta mass=2.0448, Structure: Tanimoto Coefficient=0.65, Overlap Coefficient=0.788, Fingerprint: Tanimoto Coefficient=0.632); neighbor of etiocholanolone glucuronide (delta mass=2.052, Structure: Tanimoto Coefficient=0.65, Overlap Coefficient=0.788, Fingerprint: Tanimoto Coefficient=0.632) neighbor of epiandrosterone glucuronide (delta mass=2.0084, Structure: Tanimoto Coefficient=0.737, Overlap Coefficient=0.848, Fingerprint: Tanimoto Coefficient=0.836); neighbor of etiocholanolone glucuronide (delta mass=2.0157, Structure: Tanimoto Coefficient=0.737, Overlap Coefficient=0.848, Fingerprint: Tanimoto Coefficient=0.836)
neighbor of N6,N6,N6-trimethyllysine (delta mass=15.9949, Structure: Tanimoto Coefficient=0.929, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.806)
neighbor of anthranilate (delta mass=99.0684, Structure: Tanimoto Coefficient=0.421, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.673); neighbor of kynurenine (delta mass=28.0313, Structure: Tanimoto Coefficient=0.524, Overlap Coefficient=0.733, Fingerprint: Tanimoto Coefficient=0.696); neighbor of N-formyl-L-kynurenine (delta mass=0.0364, Structure: Tanimoto Coefficient=0.478, Overlap Coefficient=0.647, Fingerprint: Tanimoto Coefficient=0.713); neighbor of N-formylanthranilate (delta mass=71.0735, Structure: Tanimoto Coefficient=0.381, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.714); neighbor of N-formylanthranilic acid (delta mass=71.0808, Structure: Tanimoto Coefficient=0.381, Overlap Coefficient=0.667, Fingerprint: Tanimoto Coefficient=0.714) neighbor of anthranilate (delta mass=99.032, Structure: Tanimoto Coefficient=0.5, Overlap Coefficient=0.9, Fingerprint: Tanimoto Coefficient=0.745); neighbor of kynurenine (delta mass=27.9949, Structure: Tanimoto Coefficient=0.882, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=0.902); neighbor of N-formyl-L-kynurenine (delta mass=0, Structure: Tanimoto Coefficient=1, Overlap Coefficient=1, Fingerprint: Tanimoto Coefficient=1); neighbor of N-formylanthranilate (delta mass=71.0371, Structure: Tanimoto Coefficient=0.611, Overlap Coefficient=0.917, Fingerprint: Tanimoto Coefficient=0.851); neighbor of N-formylanthranilic acid (delta mass=71.0444, Structure: Tanimoto Coefficient=0.611, Overlap Coefficient=0.917, Fingerprint: Tanimoto Coefficient=0.851)
neighbor of 4-vinylguaiacol sulfate (delta mass=96.0826, Structure: Tanimoto Coefficient=0.267, Overlap Coefficient=0.533, Fingerprint: Tanimoto Coefficient=0.64)
neighbor of guanidinosuccinate (delta mass=14.9308, Structure: Tanimoto Coefficient=0.353, Overlap Coefficient=0.545, Fingerprint: Tanimoto Coefficient=0.589)
neighbor of homocitrulline (delta mass=1.0025, Structure: Tanimoto Coefficient=0.444, Overlap Coefficient=0.615, Fingerprint: Tanimoto Coefficient=0.71) neighbor of homocitrulline (delta mass=1.0025, Structure: Tanimoto Coefficient=0.368, Overlap Coefficient=0.538, Fingerprint: Tanimoto Coefficient=0.681)

Table continued (p. 5/5)

Unknown	Candidate	PubChem CID	△Mass	MFTC	MSTC	MSOC
X-24766	L-3-Cyanoalanine	439742	0.0437	0.76*	0.7*	0.88
X-24798	4-hydroxy-2-nonenal	5283344	0.0074	0.23	0.15	0.91*
X-24860	Biotin	171548	0.0466	0.58*	0.41	0.75
X-24983	D-argininium(1+)	71070	0.032	0.6*	0.57	0.8

Table B.10: Evidences of automatically selected candidate molecules based on Recon

67 candidate molecules have been automatically selected for 45 unknown metabolites based on fingerprint or structural similarity to neighboring known metabolites within the Recon database. The last three columns contain the maximum values of the fingerprint Tanimoto coefficient (MFTC), structure Tanimoto coefficient (MSTC), or structure overlap coefficient (MSOC) across pairs of the candidate molecule and known neighbors of the unknown metabolite. An asterisk indicates whether values are beyond the threshold.

Evidence
neighbor of N-alpha-acetylornithine (delta mass=60.0575, Structure: Tanimoto Coefficient=0.538, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.607); neighbor of Ornithine (delta mass=18.047, Structure: Tanimoto Coefficient=0.7, Overlap Coefficient=0.875, Fingerprint: Tanimoto Coefficient=0.755)
neighbor of trans-Hexadec-2-enoyl-CoA (delta mass=847.2142, Structure: Tanimoto Coefficient=0.152, Overlap Coefficient=0.909, Fingerprint: Tanimoto Coefficient=0.229)
neighbor of heptanoylglutamine (delta mass=14.0625, Structure: Tanimoto Coefficient=0.36, Overlap Coefficient=0.562, Fingerprint: Tanimoto Coefficient=0.581)
neighbor of N-methylpipecolate (delta mass=31.0098, Structure: Tanimoto Coefficient=0.571, Overlap Coefficient=0.8, Fingerprint: Tanimoto Coefficient=0.6)

Unknown	Candidate	Pool	Database ID	∆Mass	MSTC	MSOC	MFTC
X-01911	Dihydroisomorphine-3-glucuronide	H	HMDB0060820	0.0067	0.36	1.00*	0.00
	Dihydroisomorphine-6-glucuronide	H	HMDB0061137	0.0067	0.36	1.00*	0.00
	Dihydromorphine-3-glucuronide	H	HMDB0060821	0.0067	0.36	1.00*	0.00
X-02249	Benzo[a]pyrene-9,10-oxide	R	PCID 37456	0.0351	0.47	0.77	0.64*
	Benzo[a]pyrene-4,5-oxide	R	PCID 37786	0.0351	0.48	0.85	0.64*
X-11261	octenoyl carnitine	R	PCID 53481667	0.0068	0.50	0.91*	0.61*
	2-Octenoylcarnitine	H	HMDB0013324	0.0068	0.50	0.91*	0.12
X-11334	gamma-L-Glutamyl-L-pipecolic acid	H	HMDB0038614	0.0295	0.50	1.00*	0.00
	Imidazoleacetic acid riboside	H	HMDB0002331	0.0069	0.56	1.00*	0.00
X-11357	Gamma-Aminobutyryl-lysine	H	HMDB0001959	0.0294	0.81*	1.00*	0.13
	Glycyl-Arginine	H	HMDB0028835	0.0042	0.94*	1.00*	0.13
	Isovalerylglutamic acid	H	HMDB0000726	0.0182	0.71*	0.92*	0.13
	Valyl-Asparagine	H	HMDB0029122	0.0070	0.75*	1.00*	0.13
X-11438	2-Carboxy-4-dodecanolide	H	HMDB0030987	0.0077	0.82*	1.00*	0.00
	3-Oxotetradecanoic acid	H	HMDB0010730	0.0441	0.94*	1.00*	0.00
	6-Ketomyristic acid	H	HMDB0030982	0.0441	0.94*	1.00*	0.00
X-11444	Cortolone-3-glucuronide	H	HMDB0010320	0.0075	0.51	0.95*	0.00
X-11470	Ganosporeric acid A	H	HMDB0033022	0.0139	0.48	0.90*	0.00
X-11478	Perillic acid	R	PCID 1256	0.0073	0.53	0.75	0.61*
	(4-hydroxy-3-methoxyphenyl)acetaldehyde	R	PCID 151276	0.0291	0.53	0.73	0.60*
	2-[3-(hydroxymethyl)oxiran-2-yl]phenol	H	HMDB0134037	0.0291	0.77*	0.91*	0.18
	2-hydroxy-3-(2-hydroxyphenyl)propanal	H	HMDB0134034	0.0291	0.77*	0.91*	0.18
	3-(2-Hydroxyphenyl)propanoic acid	H	HMDB0033752	0.0291	0.92*	1.00*	0.18
	3-(2,3-dihydroxyphenyl)propanal	H	HMDB0134032	0.0291	0.77*	0.91*	0.18
	3-(2,5-dihydroxyphenyl)propanal	H	HMDB0134033	0.0291	0.77*	0.91*	0.18
	3-(4-methoxyphenyl)propan-1-ol	H	HMDB0135742	0.0073	0.77*	0.91*	0.18
	3-Hydroxy-1-(4-hydroxyphenyl)-1-propanone	H	HMDB0040645	0.0291	0.77*	0.91*	0.18
	3-hydroxy-3-(2-hydroxyphenyl)propanal	H	HMDB0134035	0.0291	0.77*	0.91*	0.18
	3-hydroxy-3-phenylpropanoic acid	H	HMDB0124925	0.0291	0.92*	1.00*	0.18
	4-[3-(hydroxymethyl)oxiran-2-yl]phenol	H	HMDB0141766	0.0291	0.77*	0.91*	0.18
	D-Phenyllactic acid	H	HMDB0000563	0.0291	0.92*	1.00*	0.18
	Desaminotyrosine	H	HMDB0002199	0.0291	0.92*	1.00*	0.18
	L-3-Phenyllactic acid	H	HMDB0000748	0.0291	0.92*	1.00*	0.18
X-11491	(3a,5b,7a)-23-Carboxy-7-hydroxy-24-norcholan-3-yl-b-D-Glucopyranosiduronic acid	H	HMDB0002430	0.0059	0.68*	0.97*	0.00
	Deoxycholic acid 3-glucuronide	H	HMDB0002596	0.0059	0.68*	0.97*	0.00
X-11612	6-Thioinosinic acid	H	HMDB0060791	0.0106	0.86*	0.95*	0.00
	arabinofuranosylguanine	H	HMDB0061067	0.0310	0.74*	0.85	0.00
X-11787	L-4-hydroxyglutamic semialdehyde	R	PCID 25201126	0.0436	0.64	0.90*	0.62*
	L-lysinium(1+)	H	HMDB0062809	0.0166	0.77*	1.00*	0.13
	N-Methyl-D-aspartic acid	H	HMDB0002393	0.0436	0.69*	0.90*	0.13
X-12026	6-Methyltetrahydropterin	H	HMDB0002249	0.0433	0.73*	0.85	0.10

Table continued (p. 2 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-12026	8-Hydroxy-7-methylguanine	H	HMDB0006037	0.0069	0.73*	0.85	0.10
X-12093	L-alanyl-L-leucine	R	PCID 6992388	0.0071	0.86*	1.00*	0.94*
	Alanyl-Isoleucine	H	HMDB0028690	0.0071	0.73*	0.92*	0.20
	Alanyl-Leucine	H	HMDB0028691	0.0071	0.86*	1.00*	0.20
X-12096	Isovalerylalanine	H	HMDB0000747	0.0185	0.65	1.00*	0.13
X-12097	Isovalerylalanine	H	HMDB0000747	0.0185	0.65	1.00*	0.13
X-12100	5-hydroxy-L-tryptophan	R	PCID 439280	0.0071	0.33	0.58	0.67*
X-12112	3-(1-Pyrrolidinyl)-2-butanone	H	HMDB0039840	0.0292	0.67*	0.80	0.08
	Conhydrinone	H	HMDB0033364	0.0292	0.67*	0.80	0.08
	L-Hypoglycin A	H	HMDB0029427	0.0072	0.67*	0.80	0.08
	Physoperuvine	H	HMDB0030338	0.0292	0.67*	0.80	0.08
X-12115	Hydroxyprolyl-Glutamine	H	HMDB0028861	0.0320	0.70*	0.89	0.00
	Isoleucyl-Glutamate	H	HMDB0028906	0.0194	0.70*	0.89	0.00
	Leucyl-Glutamate	H	HMDB0028928	0.0194	0.70*	0.89	0.00
X-12117	Alanyl-Asparagine	H	HMDB0028682	0.0434	0.50	1.00*	0.13
	Asparaginyl-Alanine	H	HMDB0028724	0.0434	0.58	1.00*	0.13
	Glycyl-Gamma-glutamate	H	HMDB0028855	0.0434	0.57	1.00*	0.13
	Glycyl-Glutamine	H	HMDB0028839	0.0434	0.57	1.00*	0.13
	Glycyl-Lysine	H	HMDB0028846	0.0070	0.47	1.00*	0.13
	N-lactoyl-Leucine	H	HMDB0062176	0.0182	0.50	1.00*	0.13
X-12125	Hydroxyprolyl-Isoleucine	H	HMDB0028866	0.0182	0.61	0.92*	0.13
	Hydroxyprolyl-Leucine	H	HMDB0028867	0.0182	0.71*	1.00*	0.13
	Isoleucyl-Isoleucine	H	HMDB0028910	0.0182	0.61	0.92*	0.13
	Isoleucyl-Leucine	H	HMDB0028911	0.0182	0.71*	1.00*	0.13
	Leucyl-Isoleucine	H	HMDB0028932	0.0182	0.61	0.92*	0.13
	Leucyl-Leucine	H	HMDB0028933	0.0182	0.71*	1.00*	0.13
X-12170	adrenochrome o-semiquinone	R	PCID 10313383	0.0053	0.50	0.69	0.61*
	5-Hydroxykynurenamine	R	PCID 164719	0.0291	0.56	0.77	0.79*
	3-Hydroxykynurenamine	R	PCID 440736	0.0291	0.65	0.85	0.84*
	trans-caffeate	R	PCID 689043	0.0185	0.65	0.85	0.58*
	Isonicotinylglycine	H	HMDB0041912	0.0073	0.85*	1.00*	0.00
X-12212	(4-ethenyl-2-methoxyphenyl)oxidanesulfonic acid	H	HMDB0127980	0.0070	0.82*	0.93*	0.07
	[3-(3-oxopropyl)phenyl]oxidanesulfonic acid	H	HMDB0135299	0.0070	0.71*	0.86	0.07
	[4-(3-hydroxyprop-1-en-1-yl)phenyl]oxidanesulfonic acid	H	HMDB0135656	0.0070	0.71*	0.86	0.07
	[4-(3-oxopropyl)phenyl]oxidanesulfonic acid	H	HMDB0135301	0.0070	0.81*	0.93*	0.07
	3-[(1E)-3-hydroxyprop-1-en-1-yl]phenyloxidanesulfonic acid	H	HMDB0135303	0.0070	0.71*	0.86	0.07
	4-[(1E)-3-hydroxyprop-1-en-1-yl]phenyloxidanesulfonic acid	H	HMDB0135306	0.0070	0.71*	0.86	0.07
X-12230	(5-ethyl-2-hydroxyphenyl)oxidanesulfonic acid	H	HMDB0124986	0.0072	0.69*	0.85	0.05
	3-hydroxybenzoic acid-3-O-sulphate	H	HMDB0059968	0.0292	0.69*	0.85	0.05
	4-hydroxybenzoic acid-4-O-sulphate	H	HMDB0059982	0.0292	0.80*	0.92*	0.05
	Tyrosol 4-sulfate	H	HMDB0041785	0.0072	0.93*	1.00*	0.05

Table continued (p. 3 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-12257	[4-(4-methyl-3-oxopent-1-en-1-yl)phenyl]oxidanesulfonic acid	H	HMDB0132951	0.0106	0.61	1.00*	0.00
	5,6-dihydroxy-2-[(4-hydroxyphenyl)methylidene]-2,3-dihydro-1-benzofuran-3-one	H	HMDB0137513	0.0140	0.40	1.00*	0.00
	5,6,7-trihydroxy-2-phenyl-4H-chromen-4-one	H	HMDB0134890	0.0140	0.40	1.00*	0.00
	5,6,7-trihydroxy-3-phenyl-4H-chromen-4-one	H	HMDB0129963	0.0140	0.40	1.00*	0.00
	5,7,8-trihydroxy-2-phenyl-4H-chromen-4-one	H	HMDB0130170	0.0140	0.40	1.00*	0.00
	6-Hydroxydaidzein	H	HMDB0031715	0.0140	0.40	1.00*	0.00
	6,7-dihydroxy-2-(3-hydroxyphenyl)-4H-chromen-4-one	H	HMDB0133316	0.0140	0.40	1.00*	0.00
	6,7-dihydroxy-3-(3-hydroxyphenyl)-4H-chromen-4-one	H	HMDB0133990	0.0140	0.40	1.00*	0.00
	6,7,8-trihydroxy-3-phenyl-4H-chromen-4-one	H	HMDB0132880	0.0140	0.40	1.00*	0.00
	7-hydroxy-6-methoxy-3-phenyl-3,4-dihydro-2H-1-benzopyran-4-one	H	HMDB0132860	0.0224	0.40	1.00*	0.00
	Rhababerone	H	HMDB0034409	0.0140	0.40	1.00*	0.00
X-12263	[4-(3-hydroxybutyl)-2-methoxyphenyl]oxidanesulfonic acid	H	HMDB0135717	0.0434	0.68*	0.93*	0.12
	[4-(4-hydroxy-3-methoxyphenyl)butan-2-yl]oxysulfonic acid	H	HMDB0135718	0.0434	0.68*	0.93*	0.12
	3-[4-hydroxy-3-(sulfooxy)phenyl]-2-oxopropanoic acid	H	HMDB0125450	0.0294	0.68*	0.93*	0.12
	3-[4-methoxy-3-(sulfooxy)phenyl]propanoic acid	H	HMDB0131144	0.0070	0.68*	0.93*	0.12
	Dihydroferulic acid 4-sulfate	H	HMDB0041724	0.0070	0.78*	1.00*	0.12
X-12379	Aspartyl-Tyrosine	H	HMDB0028765	0.0303	0.50	0.92*	0.13
	Methionyl-Phenylalanine	H	HMDB0028980	0.0116	0.52	0.92*	0.13
	Phenylalanyl-Methionine	H	HMDB0029001	0.0116	0.57	0.92*	0.13
	Tyrosyl-Aspartate	H	HMDB0029101	0.0303	0.50	0.92*	0.13
X-12511	2-Hydroxyundecanoate	H	HMDB0059736	0.0196	0.79*	1.00*	0.13
	N-Acetylaminooctanoic acid	H	HMDB0059745	0.0070	0.67*	0.91*	0.13
X-12543	3-Methoxy-4-hydroxyphenylglycolaldehyde	R	PCID 440729	0.0069	0.62	0.77	0.77*
	(+/-)-2-Hydroxy-3-(2-hydroxyphenyl)propanoic acid	H	HMDB0033624	0.0069	0.86*	0.92*	0.00
	(+/-)-threo-Anethole glycol	H	HMDB0032607	0.0433	0.73*	0.85	0.00
	2-hydroxy-2-(3-methoxyphenyl)acetic acid	H	HMDB0133494	0.0069	0.73*	0.85	0.00
	2-hydroxy-3-(3-hydroxyphenyl)propanoic acid	H	HMDB0140892	0.0069	0.86*	0.92*	0.00
	3-(2,3-dihydroxyphenyl)propanoic acid	H	HMDB0134042	0.0069	0.79*	0.92*	0.00
	3-(2,4-dihydroxyphenyl)propanoic acid	H	HMDB0126386	0.0069	0.86*	0.92*	0.00
	3-(2,5-dihydroxyphenyl)propanoic acid	H	HMDB0134043	0.0069	0.79*	0.92*	0.00
	3-(3-Hydroxyphenyl)-3-hydroxypropanoic acid	H	HMDB0002643	0.0069	0.79*	0.92*	0.00
	3-(3,5-dihydroxyphenyl)propanoic acid	H	HMDB0125533	0.0069	0.92*	1.00*	0.00

Table continued (p. 4 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-12543	3-hydroxy-3-(4-hydroxyphenyl)propanoic acid	H	HMDB0124923	0.0069	0.86*	0.92*	0.00
	3-Hydroxyphenyllactate	H	HMDB0029232	0.0069	0.86*	0.92*	0.00
	3,4-Dihydroxyhydrocinnamic acid	H	HMDB0000423	0.0069	0.92*	1.00*	0.00
	Verimol J	H	HMDB0036522	0.0433	0.73*	0.85	0.00
X-12636	(2S,2'S)-Pyrosaccharopine	H	HMDB0038676	0.0295	0.59	0.92*	0.22
	gamma-L-Glutamyl-L-pipecolic acid	H	HMDB0038614	0.0295	0.67*	0.92*	0.22
X-12688	L-alanyl-L-leucine	R	PCID 6992388	0.0073	0.50	0.80	0.69*
X-12707	3,4-Dihydroxyphenylglycol O-sulfate	H	HMDB0001474	0.0070	0.67*	0.92*	0.00
X-12713	3-Methoxy-4-hydroxyphenylethyleneglycol sulfate	H	HMDB0000559	0.0070	0.72*	0.93*	0.00
	3-Methoxy-4-Hydroxyphenylglycol sulfate	H	HMDB0003332	0.0070	0.72*	0.93*	0.00
X-12818	3-hydroxy-5-methoxy-4-(sulfooxy)benzoic acid	H	HMDB0128025	0.0292	0.65	1.00*	0.00
	Methylgallic acid-O-sulphate	H	HMDB0060005	0.0292	0.65	1.00*	0.00
X-12821	6-[2-(2H-1,3-benzodioxol-5-yl)-1-carboxy-2-oxoethyl]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0129412	0.0240	0.37	1.00*	0.00
	6-[3-(2H-1,3-benzodioxol-5-yl)-3-oxopropanoyl]oxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0129411	0.0240	0.37	1.00*	0.00
X-12822	Dopaxanthin	H	HMDB0012221	0.0135	0.41	0.92*	0.10
X-12836	Epirosmanol	H	HMDB0035812	0.0218	0.38	0.91*	0.00
	Isorosmanol	H	HMDB0036661	0.0218	0.38	0.91*	0.00
X-12839	Ethylene brassylate	H	HMDB0040459	0.0315	0.75*	1.00*	0.10
X-12844	Tetrahydroaldosterone-3-glucuronide	H	HMDB0010357	0.0077	0.48	0.90*	0.14
X-12847	2-(2,3,6-trihydroxy-4-methoxyphenyl)propanoic acid	H	HMDB0125049	0.0246	0.59	0.91*	0.00
X-12860	glutaryl carnitine	R	PCID 53481622	0.0432	0.85*	1.00*	0.87*
	Arginyl-Threonine	H	HMDB0028719	0.0207	0.57	0.92*	0.22
	gamma-Glutamyllysine	H	HMDB0029154	0.0320	0.64	0.92*	0.22
	Glutamyllysine	H	HMDB0004207	0.0320	0.64	0.92*	0.22
	Hydroxyhexanoycarnitine	H	HMDB0013131	0.0068	0.95*	1.00*	0.22
X-13553	3-Methoxy-4-hydroxyphenylethyleneglycol sulfate	H	HMDB0000559	0.0070	0.72*	0.93*	0.00
	3-Methoxy-4-Hydroxyphenylglycol sulfate	H	HMDB0003332	0.0070	0.72*	0.93*	0.00
X-13684	3-mercaptolactate-cysteine disulfide	R	PCID 193536	0.0068	0.50	1.00*	0.73*
X-13728	6-amino-5[N-methylformylamino]-1-methyluracil	H	HMDB0059771	0.0071	0.80*	0.92*	0.00
X-13844	D-altro-D-manno-Heptose	H	HMDB0029952	0.0168	0.69*	0.85	0.00
	D-Glucaric acid	H	HMDB0029881	0.0196	0.80*	0.92*	0.00
	Fluorocitric acid	H	HMDB0031255	0.0396	0.69*	0.85	0.00
	Galactaric acid	H	HMDB0000639	0.0196	0.80*	0.92*	0.00
	Sedoheptulose	H	HMDB0003219	0.0168	0.69*	0.85	0.00
X-13866	N-Ribosylnicotinamide	R	PCID 439924	0.0180	0.30	0.58	0.66*
	3-(4,7-dimethoxy-2H-1,3-benzodioxol-5-yl)propanoic acid	H	HMDB0137285	0.0293	0.63	0.92*	0.10
	3-(6,7-dimethoxy-2H-1,3-benzodioxol-5-yl)propanoic acid	H	HMDB0128669	0.0293	0.63	0.92*	0.10

Table continued (p. 5 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-13866	Acetaminophen cystein	H	HMDB0060559	0.0358	0.47	1.00*	0.10
X-13874	N2-Succinyl-L-ornithine	H	HMDB0001199	0.0182	0.70*	0.88	0.10
X-14196	Histidinyl-Histidine	H	HMDB0028887	0.0294	0.45	0.91*	0.00
	Phenylbutyrylglutamine	H	HMDB0011687	0.0433	0.45	0.91*	0.00
X-14473	L,L-Cyclo(leucylprolyl)	H	HMDB0034276	0.0068	1.00*	1.00*	0.06
X-15461	(+/-)-2-Pentylthiazolidine	H	HMDB0040057	0.0111	0.67*	0.80	0.00
	5-Acetamidovalerate	H	HMDB0012175	0.0076	0.75*	1.00*	0.00
X-15469	3-hydroxyoctanoyl carnitine	H	HMDB0061634	0.0070	0.95*	1.00*	0.00
X-15497	fructoseglycine	R	PCID 3081391	0.0081	0.50	0.69	0.61*
	salsoline-1-carboxylate	R	PCID 320322	0.0071	0.58	0.85	0.61*
X-15503	dopachrome o-semiquinone	R	PCID 53481550	0.0309	0.65	0.79	0.62*
	Monoethyl phthalate	H	HMDB0002120	0.0183	0.73*	0.92*	0.00
X-15646	5-hydroxy-11-methoxy-6,8,19-trioxapentacyclo[10.7.0.0^{2,9}.0^{3,7}.0^{13,17}] nonadeca-1(12),2(9),10,13(17)-tetraene-4,16,18-trione	H	HMDB0142065	0.0327	0.46	0.92*	0.10
	N-(4-aminobutyl)-3-[3-methoxy-4-(sulfooxy)phenyl]prop-2-enimidic acid	H	HMDB0139622	0.0183	0.71*	0.94*	0.10
X-15666	Alanyl-Methionine	H	HMDB0028693	0.0035	0.80*	0.92*	0.13
	Serylaspartic acid	H	HMDB0029035	0.0222	0.69*	0.92*	0.13
X-15728	(4-ethenyl-2,6-dihydroxyphenyl)oxidanesulfonic acid	H	HMDB0128010	0.0292	0.75*	0.92*	0.07
	(4-ethyl-2-methoxyphenyl)oxidanesulfonic acid	H	HMDB0127988	0.0071	0.94*	1.00*	0.07
	2-[4-(sulfooxy)phenyl]acetic acid	H	HMDB0132500	0.0292	0.87*	1.00*	0.07
	Vanillin 4-sulfate	H	HMDB0041789	0.0292	0.82*	0.93*	0.07
X-16071	L-allysine	R	PCID 207	0.0284	0.47	0.80	0.63*
	6-amino-2-oxohexanoic acid	R	PCID 439954	0.0284	0.56	0.90*	0.59*
	2-oxoglutaramate	R	PCID 48	0.0080	0.53	0.90*	0.60
	1H-Indole-3-carboxaldehyde	H	HMDB0029737	0.0073	0.62	0.91*	0.10
	N-Butyrylglycine	H	HMDB0000808	0.0284	0.59	1.00*	0.10
X-16563	2-acetamido-5-oxopentanoate	R	PCID 192878	0.0295	0.71*	1.00*	0.83*
	1,2,5,6-Tetrahydro-4H-pyrrolo[3,2,1-ij]quinolin-4-one	H	HMDB0037113	0.0142	0.50	1.00*	0.22
	4-Amino-2-methyl-1-naphthol	H	HMDB0032887	0.0142	0.44	1.00*	0.22
	apo-[3-methylcrotonoyl-CoA:carbon-dioxide ligase (ADP-forming)]	H	HMDB0059607	0.0181	0.50	1.00*	0.22
	Isovalerylalanine	H	HMDB0000747	0.0069	0.92*	1.00*	0.22
	Isovalerylsarcosine	H	HMDB0002087	0.0069	0.92*	1.00*	0.22
X-16567	2-Hydroxydecanoate	H	HMDB0094656	0.0201	0.52	0.92*	0.22
	2-Keto-6-acetamidocaproate	H	HMDB0012150	0.0294	0.56	1.00*	0.22
	2,3-Dihydro-5-(5-methyl-2-furanyl)-1H-pyrrolizine	H	HMDB0040012	0.0142	0.41	1.00*	0.22
	Amrinone	H	HMDB0015496	0.0393	0.35	1.00*	0.22
	N-Heptanoylglycine	H	HMDB0013010	0.0069	0.92*	1.00*	0.22
	Selegiline	H	HMDB0015171	0.0222	0.39	1.00*	0.22
X-17185	(5-ethenyl-2-hydroxyphenyl)oxidanesulfonic acid	H	HMDB0124978	0.0072	0.64	1.00*	0.00

Table continued (p. 6 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-17185	Bergapten	H	HMDB0030637	0.0403	0.56	1.00	0.00
	Methoxsalen	H	HMDB0014693	0.0403	0.56	1.00	0.00
X-17299	11-Hydroxy-9-tridecenoic acid	H	HMDB0035881	0.0182	0.48	0.94*	0.00
X-17323	16-hydroxypalmitate	R	PCID 7058075	0.0422	0.32	0.58	0.64*
X-17324	5-Hydroxymethyl tolterodine	H	HMDB0013973	0.0086	0.38	0.91*	0.00
	Propafenone	H	HMDB0015313	0.0278	0.38	0.91*	0.00
X-17342	2-[4-hydroxy-3-(sulfooxy)phenyl]acetic acid	H	HMDB0125151	0.0070	0.65	0.92*	0.00
X-17367	3,4,5,6-Tetrahydrohippuric acid	H	HMDB0061679	0.0066	0.73*	0.85	0.00
X-17438	3-Hydroxydodecanedioic acid	H	HMDB0000413	0.0074	0.72*	0.93*	0.00
	Annuolide A	H	HMDB0037801	0.0137	0.68*	0.93*	0.00
X-17685	(4-ethyl-2,6-dihydroxyphenyl)oxidanesulfonic acid	H	HMDB0128030	0.0069	0.75*	1.00*	0.07
	2-hydroxy-3-(sulfooxy)benzoic acid	H	HMDB0134105	0.0295	0.75*	1.00*	0.07
	4-hydroxy-3-(sulfooxy)benzoic acid	H	HMDB0124993	0.0295	0.73*	1.00*	0.07
X-17688	Indole-3-acetic-acid-O-glucuronide	H	HMDB0060001	0.0064	0.52	1.00*	0.00
X-17689	1-(4-Hydroxy-3-methoxyphenyl)-5-(4-hydroxyphenyl)-1,4-pentadien-3-one	H	HMDB0040931	0.0220	0.41	1.00*	0.00
	2-O-Feruloyltartronic acid	H	HMDB0032957	0.0297	0.43	1.00*	0.00
	5,5',6,6'-Tetrahydroxy-3,3'-biindolyl	H	HMDB0029301	0.0032	0.41	1.00*	0.00
	6-(4-ethenylphenoxy)-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0125168	0.0067	0.43	1.00*	0.00
	7-Hydroxy-R-phenprocoumon	H	HMDB0060888	0.0220	0.41	1.00*	0.00
	Caffeoylmalic acid	H	HMDB0029318	0.0297	0.43	1.00*	0.00
	cis-Coutaric acid	H	HMDB0034620	0.0297	0.43	1.00*	0.00
	Demethoxyegonol	H	HMDB0030772	0.0220	0.41	1.00*	0.00
	Gyrocyanin	H	HMDB0034126	0.0144	0.41	1.00*	0.00
X-17699	Glutaminylalanine	H	HMDB0028790	0.0182	0.69*	0.92*	0.00
	Lysyl-Alanine	H	HMDB0028944	0.0181	0.80*	1.00*	0.00
X-18240	Alpha-Hydroxyhippuric acid	H	HMDB0002404	0.0070	0.73*	0.92*	0.00
X-18603	Vinylacetylglycine	H	HMDB0000894	0.0070	0.80*	1.00*	0.00
X-18779	1-hydroxy-4-(4-hydroxy-3-methoxyphenyl)butan-2-one	H	HMDB0133535	0.0438	0.65	0.91*	0.09
	2-(3,4-dimethoxyphenyl)propanoic acid	H	HMDB0130385	0.0438	0.62	0.91*	0.09
	2-[(4-hydroxyphenyl)methyl]propanedioic acid	H	HMDB0142178	0.0074	0.75*	0.92*	0.09
	2-benzyl-2-hydroxypropanedioic acid	H	HMDB0142180	0.0074	0.67*	0.85	0.09
	2-Methoxy-3-(4-methoxyphenyl)propanoic acid	H	HMDB0039428	0.0438	0.76*	0.92*	0.09
	3-(3-hydroxy-4-methoxyphenyl)oxirane-2-carboxylic acid	H	HMDB0140909	0.0074	0.75*	0.92*	0.09
	3-(3,5-dimethoxyphenyl)propanoic acid	H	HMDB0127493	0.0438	0.76*	0.87	0.09
	3-(4-Hydroxy-3-methoxyphenyl)-2-methylpropionic acid	H	HMDB0060737	0.0438	0.75*	0.92*	0.09
	3-(4-hydroxy-3-methoxyphenyl)oxirane-2-carboxylic acid	H	HMDB0125516	0.0074	0.75*	0.92*	0.09

Table continued (p. 7 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-18779	3-hydroxy-4-(4-hydroxy-3-methoxyphenyl)butan-2-one	H	HMDB0133478	0.0438	0.67*	0.91*	0.09
	3,4-Dihydroxyphenylvaleric acid	H	HMDB0029233	0.0438	0.62	0.91*	0.09
	4-(2,4-dihydroxy-3-methoxyphenyl)butan-2-one	H	HMDB0133526	0.0438	0.65	0.91*	0.09
	4-(2,4-dihydroxy-5-methoxyphenyl)butan-2-one	H	HMDB0133519	0.0438	0.65	0.91*	0.09
	4-(3-methylbut-2-en-1-yl)benzene-1,2,3,5-tetrol	H	HMDB0129255	0.0438	0.62	0.91*	0.09
	4-(3,4-dihydroxy-5-methoxyphenyl)butan-2-one	H	HMDB0133511	0.0438	0.76*	0.91*	0.09
	4-hydroxy-4-(4-hydroxy-3-methoxyphenyl)butan-2-one	H	HMDB0135671	0.0438	0.65	0.91*	0.09
	D-Glucaric acid	H	HMDB0029881	0.0078	0.67*	0.91*	0.09
	Galactaric acid	H	HMDB0000639	0.0078	0.67*	0.91*	0.09
	Vanilpyruvic acid	H	HMDB0011714	0.0074	0.87*	1.00*	0.09
X-18888	2-keto-3-deoxy-D-glycero-D-galactononic acid	R	PCID 22833524	0.0446	0.57	0.80	0.79*
	Seryltyrosine	H	HMDB0029051	0.0181	0.60	0.92*	0.10
X-18889	Asparaginyl-Alanine	H	HMDB0028724	0.0181	0.73*	0.92*	0.00
	Glutaminylglycine	H	HMDB0028797	0.0181	0.73*	0.92*	0.00
	Lysyl-Glycine	H	HMDB0028951	0.0183	0.73*	0.92*	0.00
	Tryptophanamide	H	HMDB0013318	0.0028	0.74*	0.93*	0.00
X-18938	3-Hydroxydodecanedioic acid	H	HMDB0000413	0.0070	0.68*	0.87	0.10
X-19561	Dihydroferuloylglycine	H	HMDB0041725	0.0072	0.63	0.92*	0.10
	N-Acetylvanilalanine	H	HMDB0011716	0.0072	0.63	0.92*	0.10
	N-lactoyl-Tyrosine	H	HMDB0062177	0.0072	0.63	0.92*	0.10
X-19913	(2Z)-2-(phenylmethylidene)octanoic acid	H	HMDB0134123	0.0475	0.36	1.00*	0.20
	(Z)-8-Decene-4,6-diyn-1-yl 3-methylbutanoate	H	HMDB0031000	0.0475	0.33	1.00*	0.20
	1,3,11(13)-Eudesmatrien-12-oic acid	H	HMDB0039011	0.0475	0.36	1.00*	0.20
	3-(1,2-dihydroxybut-3-en-1-yl)-1H-isochromen-1-one	H	HMDB0130065	0.0252	0.33	1.00*	0.20
	3-(3-ethyloxiran-2-yl)-5-hydroxy-1H-isochromen-1-one	H	HMDB0130047	0.0252	0.31	1.00*	0.20
	3-[(1E)-but-1-en-1-yl]-5,6-dihydroxy-1H-isochromen-1-one	H	HMDB0130045	0.0252	0.29	1.00*	0.20
	3-[(1E)-but-1-en-1-yl]-5,7-dihydroxy-1H-isochromen-1-one	H	HMDB0130046	0.0252	0.29	1.00*	0.20
	3-[4-hydroxy-3-(3-methylbut-2-en-1-yl)phenyl]prop-2-enoic acid	H	HMDB0135874	0.0111	0.24	1.00*	0.20
	3-hydroxy-6-[(E)-2-methoxyethenyl]-7-methyl-2H-chromen-2-one	H	HMDB0135798	0.0252	0.26	1.00*	0.20
	4-(Glutamylamino) butanoate	H	HMDB0012161	0.0071	0.38	1.00*	0.20
	5-hydroxy-3-[(1E)-3-hydroxybut-1-en-1-yl]-1H-isochromen-1-one	H	HMDB0130044	0.0252	0.29	1.00*	0.20
	5-hydroxy-3-[(1E)-4-hydroxybut-1-en-1-yl]-1H-isochromen-1-one	H	HMDB0130048	0.0252	0.29	1.00*	0.20
	5-hydroxy-6-[(E)-2-methoxyethenyl]-7-methyl-2H-chromen-2-one	H	HMDB0135797	0.0252	0.24	1.00*	0.20
	6-(3-methoxyoxiran-2-yl)-7-methyl-2H-chromen-2-one	H	HMDB0135795	0.0252	0.26	1.00*	0.20

Table continued (p. 8 / 20)

Unknown	Candidate	Pool	Database ID	$\triangle$Mass	MSTC	MSOC	MFTC
X-19913	7-(hydroxymethyl)-6-[(E)-2-methoxyethenyl]-2H-chromen-2-one	H	HMDB0135794	0.0252	0.26	1.00*	0.20
	7-hydroxy-6-(3-oxobutyl)-2H-chromen-2-one	H	HMDB0128959	0.0252	0.26	1.00*	0.20
	8-hydroxy-6-[(E)-2-methoxyethenyl]-7-methyl-2H-chromen-2-one	H	HMDB0135796	0.0252	0.26	1.00*	0.20
	Alantolactone	H	HMDB0035906	0.0475	0.36	1.00*	0.20
	Alloalantolactone	H	HMDB0036119	0.0475	0.36	1.00*	0.20
	Aspartyl-Valine	H	HMDB0028766	0.0071	0.35	1.00*	0.20
	Costunolide	H	HMDB0036688	0.0475	0.36	1.00*	0.20
	Dihydrocoriandrin	H	HMDB0033330	0.0252	0.33	1.00*	0.20
	Glycerol 1-propanoate diacetate	H	HMDB0031640	0.0041	0.32	1.00*	0.20
	Hydroxyprolyl-Threonine	H	HMDB0028873	0.0071	0.38	1.00*	0.20
	Isoalantolactone	H	HMDB0035934	0.0475	0.36	1.00*	0.20
	Isoleucyl-Threonine	H	HMDB0028917	0.0435	0.35	1.00*	0.20
	L-Acetopine	H	HMDB0039111	0.0184	0.38	1.00*	0.20
	L-N-(3-Carboxypropyl)glutamine	H	HMDB0029393	0.0071	0.35	1.00*	0.20
	Leucyl-Threonine	H	HMDB0028939	0.0435	0.35	1.00*	0.20
	Marindinin	H	HMDB0029504	0.0111	0.42	1.00*	0.20
	N2-Succinyl-L-ornithine	H	HMDB0001199	0.0071	0.35	1.00*	0.20
	Nalidixic Acid	H	HMDB0014917	0.0140	0.41	1.00*	0.20
	Pterosin E	H	HMDB0036605	0.0111	0.33	1.00*	0.20
	Salfredin B11	H	HMDB0041325	0.0252	0.29	1.00*	0.20
	Spermic acid 2	H	HMDB0013075	0.0435	0.31	1.00*	0.20
	Tetrahydrofurfuryl cinnamate	H	HMDB0036189	0.0111	0.26	1.00*	0.20
	Threoninyl-Hydroxyproline	H	HMDB0029062	0.0071	0.35	1.00*	0.20
	Threoninyl-Isoleucine	H	HMDB0029064	0.0435	0.35	1.00*	0.20
	Threoninyl-Leucine	H	HMDB0029065	0.0435	0.35	1.00*	0.20
	Valyl-Aspartate	H	HMDB0029123	0.0071	0.35	1.00*	0.20
X-21792	(2E,4E)-2,7-Dimethyl-2,4-octadienedioic acid	H	HMDB0034099	0.0065	0.36	1.00*	0.21
	2-(2,3,4-trihydroxyphenyl)propanoic acid	H	HMDB0137823	0.0299	0.36	1.00*	0.21
	2-(2,4,5-trihydroxyphenyl)propanoic acid	H	HMDB0125024	0.0299	0.36	1.00*	0.21
	2-(2,5-dihydroxy-4-methoxyphenyl)acetic acid	H	HMDB0130476	0.0299	0.36	1.00*	0.21
	2-(3,4-dihydroxy-5-methoxyphenyl)acetic acid	H	HMDB0131426	0.0299	0.36	1.00*	0.21
	2-hydroxy-2-(2-hydroxy-4-methoxyphenyl)acetic acid	H	HMDB0137136	0.0299	0.36	1.00*	0.21
	2-hydroxy-2-(4-hydroxy-3-methoxyphenyl)acetic acid	H	HMDB0133489	0.0299	0.36	1.00*	0.21
	2-hydroxy-3,4-dimethoxybenzoic acid	H	HMDB0142084	0.0299	0.36	1.00*	0.21
	2,3-Methylene suberic acid	H	HMDB0059779	0.0065	0.45	1.00*	0.21
	3-(2,4-dihydroxyphenyl)-3-hydroxypropanoic acid	H	HMDB0126478	0.0299	0.36	1.00*	0.21
	3-(2,4,5-trihydroxyphenyl)propanoic acid	H	HMDB0126489	0.0299	0.36	1.00*	0.21
	3-(3,4-Dihydroxyphenyl)lactic acid	H	HMDB0003503	0.0299	0.36	1.00*	0.21

Table continued (p. 9 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-21792	3-(3,4,5-trihydroxyphenyl)propanoic acid	H	HMDB0125529	0.0299	0.36	1.00*	0.21
	3-[(1E)-buta-1,3-dien-1-yl]-1H-isochromen-1-one	H	HMDB0130034	0.0146	0.33	1.00*	0.21
	3-hydroxy-2,4-dimethoxybenzoic acid	H	HMDB0140947	0.0299	0.36	1.00*	0.21
	3-Hydroxy-4-methoxymandelate	H	HMDB0029170	0.0299	0.36	1.00*	0.21
	3,4-Methylene suberic acid	H	HMDB0059768	0.0065	0.45	1.00*	0.21
	3,4-O-Dimethylgallic acid	H	HMDB0041662	0.0299	0.36	1.00*	0.21
	4-hydroxy-2,3-dimethoxybenzoic acid	H	HMDB0142087	0.0299	0.36	1.00*	0.21
	5-(3E-Pentenyl)tetrahydro-2-oxo-3-furancarboxylic acid	H	HMDB0030991	0.0065	0.41	1.00*	0.21
	Ethyl gallate	H	HMDB0033836	0.0299	0.36	1.00*	0.21
	Mimosine	H	HMDB0015188	0.0186	0.36	1.00*	0.21
	Syringic acid	H	HMDB0002085	0.0299	0.36	1.00*	0.21
X-21827	Ethylene brassylate	H	HMDB0040459	0.0429	0.75*	1.00*	0.00
X-21831	12-oxo-20-carboxy-leukotriene B4	R	PCID 53481457	0.0210	0.32	0.82	0.70*
	3-(1,1-Dimethyl-2-propenyl)-8-(3-methyl-2-butenyl)xanthyletin	H	HMDB0030730	0.0362	0.41	1.00*	0.00
X-22158	hexanoyl carnitine	R	PCID 6426853	0.0068	1.00*	1.00*	1.00*
	L-Hexanoylcarnitine	H	HMDB0000756	0.0068	1.00*	1.00*	0.00
X-22379	androsterone 3-glucosiduronic acid	R	PCID 114833	0.0063	1.00*	1.00*	1.00*
	5alpha-Dihydrotestosterone glucuronide	R	PCID 44263365	0.0063	0.78*	0.88	0.96*
	3-alpha-hydroxy-5-alpha-androstane-17-one 3-D-glucuronide	H	HMDB0010365	0.0063	1.00*	1.00*	0.13
X-22475	(-)-Morphine	H	HMDB0040987	0.0086	0.38	1.00*	0.22
	Arborinine	H	HMDB0030177	0.0278	0.29	1.00*	0.22
	Dihydroisomorphine	H	HMDB0060929	0.0086	0.38	1.00*	0.22
	Erysopine	H	HMDB0030256	0.0086	0.33	1.00*	0.22
	Glycylprolylhydroxyproline	H	HMDB0002171	0.0046	0.52	1.00*	0.22
	Hydromorphone	H	HMDB0014472	0.0086	0.38	1.00*	0.22
	Isothipendyl	H	HMDB0015692	0.0021	0.29	1.00*	0.22
	Letrozole	H	HMDB0015141	0.0265	0.27	1.00*	0.22
	Morphine	H	HMDB0014440	0.0086	0.38	1.00*	0.22
	N-Monodesmethyl-rizatriptan	H	HMDB0060847	0.0311	0.28	1.00*	0.22
	Norcodeine	H	HMDB0060657	0.0086	0.38	1.00*	0.22
	norhydrocodone	H	HMDB0061070	0.0086	0.38	1.00*	0.22
	Piperine	H	HMDB0029377	0.0086	0.33	1.00*	0.22
	Probenecid	H	HMDB0015166	0.0244	0.32	1.00*	0.22
	Secodemethylclausenamide	H	HMDB0032961	0.0086	0.38	1.00*	0.22
X-22836	L-allysine	R	PCID 207	0.0073	0.62	0.80	0.64*
	6-amino-2-oxohexanoic acid	R	PCID 439954	0.0073	0.50	0.70	0.59*
	L-cis-4-(Hydroxymethyl)-2-pyrrolidinecarboxylic acid	H	HMDB0029425	0.0073	0.75*	0.90*	0.10
	L-trans-5-Hydroxy-2-piperidinecarboxylic acid	H	HMDB0029426	0.0073	0.75*	0.90*	0.10
X-23196	gama-L-glutamyl-L-alanine	R	PCID 11183554	0.0177	0.60	0.80	0.94*
	L-leucyl-L-serine	R	PCID 40489070	0.0187	0.60	0.80	0.73*
	5-L-Glutamyl-L-alanine	R	PCID 440103	0.0177	0.60	0.80	0.94*

Table continued (p. 10/20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-23196	N-acetylserotonin	R	PCID 903	0.0025	0.48	0.69	0.79*
	2-Hydroxydecanedioic acid	H	HMDB0000424	0.0074	0.88*	0.93*	0.10
X-23291	3-ethylphenyl Sulfate	H	HMDB0062721	0.0074	0.79*	0.92*	0.07
	4-ethylphenylsulfate	H	HMDB0062551	0.0074	0.79*	0.92*	0.07
X-23304	3,4,5-trihydroxy-6-(4-hydroxybenzoyloxy)oxane-2-carboxylic acid	H	HMDB0125175	0.0071	0.45	1.00*	0.00
	6-(4-carboxyphenoxy)-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0125174	0.0071	0.45	1.00*	0.00
X-23306	(R) 2,3-Dihydroxy-3-methylvalerate	H	HMDB0012140	0.0075	0.90*	1.00*	0.00
	(R)-mevalonate	H	HMDB0059629	0.0075	0.73*	0.89	0.00
	D-Xylono-1,5-lactone	H	HMDB0011676	0.0289	0.73*	0.89	0.00
X-23423	2-Amino-3-oxoadipate	R	PCID 440714	0.0289	0.56	0.75	0.65*
	L-argininium(1+)	H	HMDB0062762	0.0425	0.71*	0.83	0.20
X-23429	2-Hydroxy-2-(2-oxopropyl)butanedioic acid	H	HMDB0059927	0.0291	0.79*	0.92*	0.00
	3-Dehydroquinate	H	HMDB0012710	0.0291	0.67*	0.83	0.00
X-23518	4-methoxy-3-(sulfooxy)benzoic acid	H	HMDB0140929	0.0070	0.65	0.92*	0.10
	Vanillic acid 4-sulfate	H	HMDB0041788	0.0070	0.75*	1.00*	0.10
X-23581	1-piperideine-6-carboxylate	R	PCID 45266761	0.0073	0.67*	0.89	0.57
	(+/-)-4-Methylene-2-pyrrolidinecarboxylic acid	H	HMDB0029434	0.0073	0.80*	0.89	0.10
	(S)-2,3,4,5-tetrahydropyridine-2-carboxylate	H	HMDB0059657	0.0073	0.67*	0.89	0.10
X-23583	acetamidopropanal	R	PCID 5460495	0.0074	0.45	0.62	0.62*
	D-proline	R	PCID 8988	0.0074	0.60	0.75	0.61*
	4-Amino-2-methylenebutanoic acid	H	HMDB0030409	0.0074	0.78*	0.88	0.08
X-23590	Glycyl-Glutamate	H	HMDB0028840	0.0196	0.73*	0.92*	0.00
	Glycyl-Glutamine	H	HMDB0028839	0.0042	0.86*	1.00*	0.00
	Glycyl-Lysine	H	HMDB0028846	0.0406	0.73*	0.92*	0.00
	N-lactoyl-Leucine	H	HMDB0062176	0.0294	0.73*	0.92*	0.00
X-23593	N-(omega)-Hydroxyarginine	R	PCID 440849	0.0043	0.43	0.69	0.61*
	Alanyl-Threonine	H	HMDB0028697	0.0069	0.67*	0.92*	0.00
	N-(gamma-Glutamyl)ethanolamine	H	HMDB0039222	0.0069	0.76*	1.00*	0.00
X-23644	1H-Indole-3-carboxaldehyde	H	HMDB0029737	0.0281	0.67*	0.89	0.00
	2,4-Dimethyl-1H-indole	H	HMDB0032967	0.0082	0.67*	0.89	0.00
X-23647	L-Prolinylglycine	R	PCID 6426709	0.0294	0.75*	1.00*	0.74*
	Glycylproline	R	PCID 79101	0.0294	0.75*	1.00*	0.79*
X-23648	Glutaminylaspartic acid	H	HMDB0028793	0.0068	0.70*	0.88	0.00
	Lysyl-Aspartate	H	HMDB0028947	0.0296	0.70*	0.88	0.00
X-23747	N1,N8-diacetylspermidine	R	PCID 389613	0.0074	0.71*	1.00*	0.98*
X-23776	Allysine	R	PCID 207	0.0435	0.67*	0.80	0.08
	2-Keto-6-aminocaproate	R	PCID 439954	0.0435	0.67*	0.80	0.08
	(S)-2-amino-6-oxohexanoate	H	HMDB0059595	0.0435	0.67*	0.80	0.08
	(S)-5-Amino-3-oxohexanoate	H	HMDB0012131	0.0435	0.67*	0.80	0.08
	2-Aminoheptanoate	H	HMDB0094649	0.0071	0.67*	0.80	0.08
	L-cis-4-(Hydroxymethyl)-2-pyrrolidinecarboxylic acid	H	HMDB0029425	0.0435	0.67*	0.80	0.08

Table continued (p. 11/20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-23776	L-trans-5-Hydroxy-2-piperidinecarboxylic acid	H	HMDB0029426	0.0435	0.82*	0.90*	0.08
	N-(2-Carboxymethyl)-morpholine	H	HMDB0061156	0.0435	0.67*	0.80	0.08
	N-Butyrylglycine	H	HMDB0000808	0.0435	0.67*	0.80	0.08
X-23908	1-hydroxyhexanoylglycine	H	HMDB0094718	0.0072	0.57	1.00*	0.00
	Asparaginyl-Glycine	H	HMDB0028731	0.0179	0.57	1.00*	0.00
	Glutarylglycine	H	HMDB0000590	0.0292	0.57	1.00*	0.00
	Glycyl-Asparagine	H	HMDB0028836	0.0179	0.69*	1.00*	0.00
	N-lactoyl-Valine	H	HMDB0062181	0.0072	0.69*	1.00*	0.00
X-23997	6-ethoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0131286	5.0e-4	0.75*	1.00*	0.00
	Isopropyl beta-D-glucoside	H	HMDB0032705	0.0358	0.67*	0.93*	0.00
X-24270	(1R,2S,5R,6S)-6-(3,4-Dihydroxyphenyl)-2-(3,4-methylenedioxyphenyl)-3,7-dioxabicyclo-[3,3,0]octane	H	HMDB0059746	0.0230	0.42	0.92*	0.00
	1,5,8-Trihydroxy-3-methyl-2-prenylxanthone	H	HMDB0034182	0.0230	0.42	0.92*	0.00
	15,16-dihydroxy-11-methoxy-6,8,20-trioxapentacyclo[10.8.0.0^{2},^{9}.0^{3},^{7}.0^{14},^{19}]icosa-1(12),2(9),10,14(19),15,17-hexaen-13-one	H	HMDB0128882	0.0133	0.42	0.92*	0.00
	3,4,5-trihydroxy-6-[3-(3-hydroxyphenyl)propanoyl]oxyoxane-2-carboxylic acid	H	HMDB0124983	0.0078	0.44	0.92*	0.00
	3,4,5-trihydroxy-6-[3-(4-hydroxyphenyl)propanoyl]oxyoxane-2-carboxylic acid	H	HMDB0125170	0.0078	0.44	0.92*	0.00
	6-[3-(2-carboxyethyl)phenoxy]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0124982	0.0078	0.44	0.92*	0.00
	6-[4-(2-carboxyethyl)phenoxy]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0125169	0.0078	0.44	0.92*	0.00
X-24328	(S)-carnitinium	H	HMDB0062634	0.0313	0.65	1.00*	0.00
X-24330	Porphobilinogen	R	PCID 1021	0.0187	0.42	1.00*	0.42
	2-(2-hydroxy-3,4-dimethoxyphenyl)propanoic acid	H	HMDB0133894	0.0074	0.31	1.00*	0.21
	2-Phenylethyl benzoate	H	HMDB0033946	0.0227	0.29	1.00*	0.21
	3-(2-hydroxy-3,4-dimethoxyphenyl)propanoic acid	H	HMDB0142080	0.0074	0.35	1.00*	0.21
	3-(3,4-dihydroxy-5-methoxyphenyl)-2-oxopropanoic acid	H	HMDB0126510	0.0290	0.35	1.00*	0.21
	3-(3,4-dihydroxy-5-methoxyphenyl)oxirane-2-carboxylic acid	H	HMDB0125587	0.0290	0.35	1.00*	0.21
	3-(4-Hydroxy-3-methoxyphenyl)-2-methyllactic acid	H	HMDB0060736	0.0074	0.35	1.00*	0.21
	3-Nitrotyrosine	H	HMDB0001904	0.0177	0.35	1.00*	0.21
	3,4-Dimethoxy-1,2-benzenedicarboxylic acid	H	HMDB0033842	0.0290	0.31	1.00*	0.21
	3,4-Methylenesebacic acid	H	HMDB0059729	0.0438	0.55	1.00*	0.21
	4-Hydroxy-(3',4'-dihydroxyphenyl)-valeric acid	H	HMDB0041679	0.0074	0.42	1.00*	0.21
	4,5-Dimethoxy-1,2-benzenedicarboxylic acid	H	HMDB0033893	0.0290	0.31	1.00*	0.21

Table continued (p. 12 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-24330	5-(3,4,5-trihydroxyphenyl)pentanoic acid	H	HMDB0141549	0.0074	0.42	1.00*	0.21
	5-(3,5-dihydroxyphenyl)-4-hydroxypentanoic acid	H	HMDB0141547	0.0074	0.42	1.00*	0.21
	Alanyl-Histidine	H	HMDB0028689	0.0299	0.32	1.00*	0.21
	Chorismate	H	HMDB0012199	0.0290	0.31	1.00*	0.21
	Dihydrosinapic acid	H	HMDB0041727	0.0074	0.35	1.00*	0.21
	Genipin	H	HMDB0035830	0.0074	0.59	1.00*	0.21
	Histidinyl-Alanine	H	HMDB0028878	0.0299	0.32	1.00*	0.21
	methyl 2-hydroxy-3-(4-hydroxy-3-methoxyphenyl)propanoate	H	HMDB0124956	0.0074	0.35	1.00*	0.21
	p-Tolyl phenylacetate	H	HMDB0041564	0.0227	0.29	1.00*	0.21
	Phenylmethyl benzeneacetate	H	HMDB0033251	0.0227	0.29	1.00*	0.21
	Prephenate	H	HMDB0012283	0.0290	0.42	1.00*	0.21
X-24339	Xanthosine 5'-phosphate	R	PCID 73323	0.0336	0.73*	0.92*	0.94*
	3-(3,4-dihydroxy-5-methoxyphenyl)-1-(2,4,6-trihydroxy-3-methoxyphenyl)propane-1,2-dione	H	HMDB0128763	0.0038	0.44	0.92*	0.20
	5-formamido-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide(2-)	H	HMDB0062708	0.0325	0.73*	0.92*	0.20
	Xanthylic acid	H	HMDB0001554	0.0336	0.73*	0.92*	0.20
X-24343	(4-ethenyl-2,6-dihydroxyphenyl)oxidanesulfonic acid	H	HMDB0128010	0.0289	0.88*	0.93*	0.07
	(4-ethyl-2-methoxyphenyl)oxidanesulfonic acid	H	HMDB0127988	0.0074	0.88*	0.93*	0.07
	2-[4-(sulfooxy)phenyl]acetic acid	H	HMDB0132500	0.0289	0.67*	0.80	0.07
	Vanillin 4-sulfate	H	HMDB0041789	0.0289	0.88*	0.93*	0.07
X-24344	(5-ethyl-2-hydroxyphenyl)oxidanesulfonic acid	H	HMDB0124986	0.0075	0.73*	0.92*	0.04
	3-hydroxybenzoic acid-3-O-sulphate	H	HMDB0059968	0.0289	0.73*	0.92*	0.04
	4-hydroxybenzoic acid-4-O-sulphate	H	HMDB0059982	0.0289	0.73*	0.92*	0.04
	Tyrosol 4-sulfate	H	HMDB0041785	0.0075	0.73*	0.92*	0.04
X-24353	Homofukinolide	H	HMDB0034659	0.0229	0.34	0.92*	0.00
X-24357	11b,17a,21-Trihydroxypreg-nenolone	H	HMDB0006760	0.0075	0.79*	0.92*	0.13
	11b,21-Dihydroxy-3,20-oxo-5b-pregnan-18-al	H	HMDB0006754	0.0075	0.73*	0.88	0.13
	18-Hydroxy-11-dehydrotetrahydrocorticosterone	H	HMDB0013157	0.0075	0.79*	0.88	0.13
	3a,11b,21-Trihydroxy-20-oxo-5b-pregnan-18-al	H	HMDB0006753	0.0075	0.73*	0.88	0.13
	Dihydrocortisol	H	HMDB0003259	0.0075	0.86*	0.96*	0.13
X-24358	19-Noraldosterone	H	HMDB0041795	0.0230	0.59	0.94*	0.00
	Diosbulbin E	H	HMDB0036779	0.0134	0.53	0.93*	0.00
	Diosbulbin G	H	HMDB0036782	0.0134	0.53	0.93*	0.00
	Gibberellin A113	H	HMDB0032819	0.0230	0.60	1.00*	0.00
	Gibberellin A37	H	HMDB0035047	0.0230	0.56	1.00*	0.00
	Gibberellin A44	H	HMDB0036901	0.0230	0.56	1.00*	0.00
	Gibberellin A64	H	HMDB0036894	0.0230	0.60	1.00*	0.00
	Gibberellin A94	H	HMDB0031365	0.0134	0.50	0.93*	0.00

Table continued (p. 13/20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-24361	testosterone 3-glucosiduronic acid	R	PCID 108192	0.0073	0.74*	0.85	0.84*
	16-Glucuronide-estriol	R	PCID 122281	0.0291	0.65	0.79	0.63*
	15-Hydroxynorandrostene-3,17-dione glucuronide	H	HMDB0010353	0.0291	0.69*	0.82	0.00
	Dehydroepiandrosterone 3-glucuronide	H	HMDB0010348	0.0073	0.94*	0.97*	0.00
	Dehydroisoandrosterone 3-glucuronide	H	HMDB0010327	0.0073	0.94*	0.97*	0.00
X-24400	6-Methyltetrahydropterin	H	HMDB0002249	0.0442	0.73*	0.85	0.10
	8-Hydroxy-7-methylguanine	H	HMDB0006037	0.0078	0.73*	0.85	0.10
X-24402	2-(4-Methyl-5-thiazolyl)ethyl formate	H	HMDB0032420	0.0464	0.40	1.00*	0.22
	2,5-Dihydro-4,5-dimethyl-2-(1-methylpropyl)thiazole	H	HMDB0037514	0.0264	0.38	1.00*	0.22
	2,5-Dihydro-4,5-dimethyl-2-(2-methylpropyl)thiazole	H	HMDB0037159	0.0264	0.38	1.00*	0.22
	8-(Methylthio)octanenitrile	H	HMDB0038438	0.0264	0.35	1.00*	0.22
	Rasagiline	H	HMDB0015454	0.0230	0.50	1.00*	0.22
	Tetrahydrodipicolinate	H	HMDB0012289	0.0286	0.50	1.00*	0.22
	Valero-1,5-lactam	H	HMDB0061932	0.0261	0.54	1.00*	0.22
X-24410	acetamidopropanal	R	PCID 5460495	0.0072	0.67*	0.86	0.67*
	D-proline	R	PCID 8988	0.0072	0.70*	0.88	0.62*
	Acetamidopropanal	H	HMDB0012880	0.0072	0.67*	0.86	0.00
	Caproate (6:0)	H	HMDB0061883	0.0054	0.70*	0.88	0.00
	N-(2-Methylpropyl)acetamide	H	HMDB0034203	0.0292	0.67*	0.86	0.00
	N-(3-Methylbutyl)acetamide	H	HMDB0031651	0.0292	0.67*	0.86	0.00
X-24411	acetamidopropanal	R	PCID 5460495	0.0072	0.50	0.75	0.63*
	D-proline	R	PCID 8988	0.0072	0.64	0.88	0.62*
X-24417	testosterone 3-glucosiduronic acid	R	PCID 108192	0.0077	0.74*	0.85	0.84*
	16-Glucuronide-estriol	R	PCID 122281	0.0287	0.65	0.79	0.63*
	15-Hydroxynorandrostene-3,17-dione glucuronide	H	HMDB0010353	0.0287	0.69*	0.82	0.13
	Dehydroepiandrosterone 3-glucuronide	H	HMDB0010348	0.0077	0.94*	0.97*	0.13
	Dehydroisoandrosterone 3-glucuronide	H	HMDB0010327	0.0077	0.94*	0.97*	0.13
X-24452	3-Hydroxy-N6,N6,N6-trimethyl-L-lysine	R	PCID 439460	0.0071	0.93*	1.00*	0.81*
	N6-Acetyl-5S-hydroxy-L-lysine	H	HMDB0033891	0.0435	0.69*	0.85	0.00
X-24455	N-acetyl-5-methoxykynuramine	R	PCID 390658	0.0294	0.52	0.80	0.71*
	N-formyl-L-kynurenine	R	PCID 910	0.0070	1.00*	1.00*	1.00*
	L-Formylkynurenine	H	HMDB0060485	0.0070	1.00*	1.00*	0.19
	N'-Formylkynurenine	H	HMDB0001200	0.0070	1.00*	1.00*	0.19
X-24462	N-Methylcalystegine C1	H	HMDB0036394	0.0292	0.73*	1.00*	0.06
X-24490	2-hydroxy-3-[4-hydroxy-3-methoxy-5-(3-methylbut-2-en-1-yl)phenyl]propanoic acid	H	HMDB0133288	0.0180	0.48	0.91*	0.00
	Valyl-Tyrosine	H	HMDB0029139	0.0292	0.48	0.91*	0.00
X-24494	11-Oxo-androsterone glucuronide	H	HMDB0010338	0.0082	0.97*	1.00*	0.13
X-24498	cis-beta-D-Glucosyl-2-hydroxycinnamate	R	PCID 5316113	0.0078	0.27	0.53	0.64*
X-24514	putreanine	R	PCID 53477800	0.0440	0.35	0.55	0.59*
X-24518	Alanyl-Asparagine	H	HMDB0028682	0.0434	0.50	1.00*	0.13
	Asparaginyl-Alanine	H	HMDB0028724	0.0434	0.58	1.00*	0.13

Table continued (p. 14 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-24518	Glycyl-Gamma-glutamate	H	HMDB0028855	0.0434	0.57	1.00*	0.13
	Glycyl-Glutamine	H	HMDB0028839	0.0434	0.57	1.00*	0.13
	Glycyl-Lysine	H	HMDB0028846	0.0070	0.47	1.00*	0.13
	N-lactoyl-Leucine	H	HMDB0062176	0.0182	0.50	1.00*	0.13
X-24542	Vanilloylglycine	H	HMDB0060026	0.0071	0.65	0.92*	0.00
X-24588	3,4,5-trihydroxy-6-3-hydroxy-4-[(E)-2-(5-hydroxy-2,2-dimethyl-2H-chromen-7-yl)ethenyl]phenoxyoxane-2-carboxylic acid	H	HMDB0134169	0.0146	0.33	1.00*	0.21
	3,4,5-trihydroxy-6-5-hydroxy-2-[(E)-2-(5-hydroxy-2,2-dimethyl-2H-chromen-7-yl)ethenyl]phenoxyoxane-2-carboxylic acid	H	HMDB0134170	0.0146	0.33	1.00*	0.21
	6-(7-[(E)-2-(2,4-dihydroxyphenyl)ethenyl]-2,2-dimethyl-2H-chromen-5-yloxy)-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0134168	0.0146	0.33	1.00*	0.21
	6-4-[(E)-2-3,5-dihydroxy-4-[(1E)-3-methylbuta-1,3-dien-1-yl]phenylethenyl]-2-hydroxyphenoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0129087	0.0146	0.33	1.00*	0.21
	6-5-[(E)-2-(3,4-dihydroxyphenyl)ethenyl]-3-hydroxy-2-[(1E)-3-methylbuta-1,3-dien-1-yl]phenoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0129086	0.0146	0.33	1.00*	0.21
	6-5-[(E)-2-3,5-dihydroxy-4-[(1E)-3-methylbuta-1,3-dien-1-yl]phenylethenyl]-2-hydroxyphenoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0129088	0.0146	0.33	1.00*	0.21
	Glucosylgalactosyl hydroxylysine	H	HMDB0000585	0.0389	0.19	1.00*	0.21
X-24736	Glycylleucine	R	PCID 92843	0.0184	0.50	0.73	0.71*
	Leucylglycine	R	PCID 97364	0.0184	0.41	0.64	0.68*
X-24738	4-Hydroxystachydrine	H	HMDB0029230	0.0072	0.91*	1.00*	0.00
	Betonicine	H	HMDB0029412	0.0072	0.91*	1.00*	0.00
	Turicine	H	HMDB0029409	0.0072	0.91*	1.00*	0.00
X-24766	L-3-Cyanoalanine	R	PCID 439742	0.0437	0.70*	0.88	0.76*
	3-Amino-2-piperidone	H	HMDB0000323	0.0073	0.70*	0.88	0.00
	epsilon-Caprolactone	H	HMDB0060476	0.0185	0.70*	0.88	0.00
	L-3-Cyanoalanine	H	HMDB0060245	0.0437	0.70*	0.88	0.00
X-24796	L-2,3-Dihydrodipicolinate	H	HMDB0012247	0.0400	0.50	1.00*	0.13
X-24798	4-Hydroxynonenal	R	PCID 5283344	0.0074	0.15	0.91*	0.23
	(E)-3-decen-1-ol	H	HMDB0013810	0.0438	0.15	0.91*	0.21
	1-Decen-3-ol	H	HMDB0039854	0.0438	0.15	0.91*	0.21
	2-Decanone	H	HMDB0031409	0.0438	0.15	0.91*	0.21
	2-DECENOL	H	HMDB0032532	0.0438	0.15	0.91*	0.21
	2-Nonenoic acid	H	HMDB0031271	0.0074	0.15	0.91*	0.21
	3-Decanone	H	HMDB0032212	0.0438	0.15	0.91*	0.21
	3,4-Epoxynonanal	H	HMDB0060286	0.0074	0.15	0.91*	0.21
	4-Oxononanal	H	HMDB0034716	0.0074	0.15	0.91*	0.21
	6-Butyltetrahydro-2H-pyran-2-one	H	HMDB0031514	0.0074	0.15	0.91*	0.21

Table continued (p. 15 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-24798	9-Decenol	H	HMDB0059861	0.0438	0.15	0.91*	0.21
	cis-4-Decenol	H	HMDB0032206	0.0438	0.17	0.91*	0.21
	Decanal	H	HMDB0011623	0.0438	0.17	1.00*	0.21
	Dihydro-5-pentyl-2(3H)-furanone	H	HMDB0031531	0.0074	0.15	0.91*	0.21
X-24799	(+/-)-2-Hydroxy-3-(2-hydroxyphenyl)propanoic acid	H	HMDB0033624	0.0289	0.38	1.00*	0.21
	2-(2,4-dihydroxyphenyl)propanoic acid	H	HMDB0133788	0.0289	0.38	1.00*	0.21
	2-(3-hydroxy-5-methoxyphenyl)acetic acid	H	HMDB0131428	0.0289	0.38	1.00*	0.21
	2-(3,4-dihydroxyphenyl)propanoic acid	H	HMDB0129348	0.0289	0.38	1.00*	0.21
	2-hydroxy-2-(3-methoxyphenyl)acetic acid	H	HMDB0133494	0.0289	0.38	1.00*	0.21
	2-hydroxy-2-(4-methoxyphenyl)acetic acid	H	HMDB0140294	0.0289	0.38	1.00*	0.21
	2-hydroxy-3-(3-hydroxyphenyl)propanoic acid	H	HMDB0140892	0.0289	0.38	1.00*	0.21
	2,6-Dimethoxybenzoic acid	H	HMDB0029273	0.0289	0.38	1.00*	0.21
	3-(2,3-dihydroxyphenyl)propanoic acid	H	HMDB0134042	0.0289	0.38	1.00*	0.21
	3-(2,4-dihydroxyphenyl)propanoic acid	H	HMDB0126386	0.0289	0.38	1.00*	0.21
	3-(2,5-dihydroxyphenyl)propanoic acid	H	HMDB0134043	0.0289	0.38	1.00*	0.21
	3-(3-Hydroxyphenyl)-3-hydroxypropanoic acid	H	HMDB0002643	0.0289	0.38	1.00*	0.21
	3-(3,5-dihydroxyphenyl)propanoic acid	H	HMDB0125533	0.0289	0.38	1.00*	0.21
	3-hydroxy-3-(4-hydroxyphenyl)propanoic acid	H	HMDB0124923	0.0289	0.38	1.00*	0.21
	3-Hydroxyphenyllactate	H	HMDB0029232	0.0289	0.38	1.00*	0.21
	3,4-Dihydroxyhydrocinnamic acid	H	HMDB0000423	0.0289	0.38	1.00*	0.21
	3,4-Dimethoxybenzoic acid	H	HMDB0059763	0.0289	0.38	1.00*	0.21
	3,5-dimethoxybenzoic acid	H	HMDB0127495	0.0289	0.38	1.00*	0.21
	4-Hydroxy-4-methyl-7-decenoic acid gamma-lactone	H	HMDB0036191	0.0439	0.42	1.00*	0.21
	5-Pentyl-3h-furan-2-one	H	HMDB0032463	0.0439	0.38	1.00*	0.21
	6-(3-Hexenyl)tetrahydro-2H-pyran-2-one	H	HMDB0037829	0.0439	0.41	1.00*	0.21
	Allyl cyclohexylacetate	H	HMDB0040595	0.0439	0.41	1.00*	0.21
	cis-3-Hexenyl tiglate	H	HMDB0038279	0.0439	0.38	1.00*	0.21
	Furfuryl isovalerate	H	HMDB0039874	0.0075	0.38	1.00*	0.21
	Furfuryl pentanoate	H	HMDB0037727	0.0075	0.38	1.00*	0.21
	Isohomovanillic acid	H	HMDB0000333	0.0289	0.38	1.00*	0.21
	Maltol propionate	H	HMDB0037274	0.0289	0.38	1.00*	0.21
	Meta-hydroxyphenylhydracrylic Acid	H	HMDB0062595	0.0289	0.38	1.00*	0.21
	Methyl 4,8-decadienoate	H	HMDB0034708	0.0439	0.38	1.00*	0.21
	Peperinic acid	H	HMDB0038181	0.0075	0.38	1.00*	0.21
X-24807	Bethanechol	H	HMDB0015154	0.0054	0.64	1.00*	0.00
	Malonyl-Carnitin	H	HMDB0062496	0.0184	0.64	1.00*	0.00
X-24838	(2E)-1-3-[(3,3-dimethyloxiran-2-yl)methyl]-2,4,6-trihydroxyphenyl-3-phenylprop-2-en-1-one	H	HMDB0129262	0.0017	0.22	1.00*	0.13
	(2E)-3-(3-hydroxyphenyl)-1-[2,4,6-trihydroxy-3-(3-methylbut-2-en-1-yl)phenyl]prop-2-en-1-one	H	HMDB0129266	0.0017	0.15	1.00*	0.13

Table continued (p. 16 / 20)

Unknown	Candidate	Pool	Database ID	$\triangle$Mass	MSTC	MSOC	MFTC
X-24838	(2E)-3-phenyl-1-[2,3,4,6-tetrahydroxy-5-(3-methylbut-2-en-1-yl)phenyl]prop-2-en-1-one	H	HMDB0129264	0.0017	0.15	1.00*	0.13
	(2E)-3-phenyl-1-[2,4,6-trihydroxy-3-(4-hydroxy-3-methylbut-2-en-1-yl)phenyl]prop-2-en-1-one	H	HMDB0129258	0.0017	0.15	1.00*	0.13
	(2E)-3-phenyl-1-2,4,6-trihydroxy-3-[(2E)-4-hydroxy-3-methylbut-2-en-1-yl]phenylprop-2-en-1-one	H	HMDB0129260	0.0017	0.15	1.00*	0.13
	(S)-4',5,7-Trihydroxy-3'-prenylflavanone	H	HMDB0029866	0.0017	0.22	1.00*	0.13
	(S)-4',5,7-Trihydroxy-6-prenylflavanone	H	HMDB0037247	0.0017	0.22	1.00*	0.13
	1-(2,4-dihydroxyphenyl)-3-[4-hydroxy-3-(4-hydroxy-3-methylbut-2-en-1-yl)phenyl]prop-2-en-1-one	H	HMDB0135875	0.0017	0.15	1.00*	0.13
	1-(2,4-dihydroxyphenyl)-3-3-[(3,3-dimethyloxiran-2-yl)methyl]-4-hydroxyphenylprop-2-en-1-one	H	HMDB0135878	0.0017	0.18	1.00*	0.13
	1-(2,4-dihydroxyphenyl)-3-4-hydroxy-3-[(2E)-4-hydroxy-3-methylbut-2-en-1-yl]phenylprop-2-en-1-one	H	HMDB0135876	0.0017	0.15	1.00*	0.13
	1-(2H-1,3-benzodioxol-5-yl)-3-hydroxy-3-(4-methoxy-1-benzofuran-5-yl)propan-1-one	H	HMDB0129367	0.0347	0.22	1.00*	0.13
	1-(3,4-dihydroxyphenyl)-7-(3-methoxyphenyl)hept-1-ene-3,5-dione	H	HMDB0133503	0.0017	0.26	1.00*	0.13
	1-(4-hydroxy-3-methoxyphenyl)-3-(4-methoxy-1-benzofuran-5-yl)propane-1,3-dione	H	HMDB0129392	0.0347	0.22	1.00*	0.13
	1-[2,4-dihydroxy-6-methoxy-3-(3-methylbut-2-en-1-yl)phenyl]-3-phenylpropan-1-one	H	HMDB0124921	0.0381	0.26	1.00*	0.13
	1-Hydroxy-3,5-dimethoxy-2-prenylxanthone	H	HMDB0041251	0.0017	0.18	1.00*	0.13
	1,7-bis(3-methoxyphenyl)heptane-3,5-dione	H	HMDB0133470	0.0381	0.34	1.00*	0.13
	11-Hydroxytubotaiwine	H	HMDB0040799	0.0493	0.34	1.00*	0.13
	15-(3-methylbut-2-en-1-yl)-8,17-dioxatetracyclo[8.7.0.0^{2},^{7}.0^{11},^{16}] heptadeca-2,4,6,11(16),12,14-hexaene-3,5,14-triol	H	HMDB0140717	0.0017	0.22	1.00*	0.13
	15-(3-methylbut-2-en-1-yl)-8,17-dioxatetracyclo[8.7.0.0^{2},^{7}.0^{11},^{16}] heptadeca-2(7),3,5,11(16),12,14-hexaene-4,5,14-triol	H	HMDB0140711	0.0017	0.18	1.00*	0.13
	15-(3-methylbut-2-en-1-yl)-8,17-dioxatetracyclo[8.7.0.0^{2},^{7}.0^{11},^{16}] heptadeca-2(7),3,5,11(16),12,14-hexaene-5,13,14-triol	H	HMDB0140712	0.0017	0.18	1.00*	0.13
	15-(3-methylbut-2-en-1-yl)-8,17-dioxatetracyclo[8.7.0.0^{2},^{7}.0^{11},^{16}] heptadeca-2(7),3,5,11(16),12,14-hexaene-5,6,14-triol	H	HMDB0140716	0.0017	0.18	1.00*	0.13
	15-(4-hydroxy-3-methylbut-2-en-1-yl) -8,17-dioxatetracyclo[8.7.0.0^{2},^{7}.0^{11},^{16}] heptadeca-2(7),3,5,11(16),12,14-hexaene-5,14-diol	H	HMDB0140713	0.0017	0.18	1.00*	0.13

Table continued (p. 17/20)

Unknown	Candidate	Pool	Database ID	$\triangle$Mass	MSTC	MSOC	MFTC
X-24838	15-[(3,3-dimethyloxiran-2-yl)methyl] -8,17-dioxatetracyclo[8.7.0.0^{2},^{7}.0^{11},^{16}] heptadeca-2(7),3,5,11(16),12,14-hexaene-5,14-diol	H	HMDB0140715	0.0017	0.22	1.00*	0.13
	15-Deacetylneosolaniol	H	HMDB0036157	0.0228	0.36	1.00*	0.13
	2-O-(4-O-Methyl-a-D-glucopyranuronosyl)-D-xylose	H	HMDB0029917	0.0288	0.28	1.00*	0.13
	2-O-a-D-Galactopyranuronosyl-L-rhamnose	H	HMDB0040159	0.0288	0.28	1.00*	0.13
	2,2-Bis[4-(2,3-epoxypropoxy)phenyl]propane	H	HMDB0032737	0.0381	0.15	1.00*	0.13
	2',4'-Dihydroxy-7-methoxy-8-prenylflavan	H	HMDB0036401	0.0381	0.22	1.00*	0.13
	2',7-Dihydroxy-4'-methoxy-8-prenylflavan	H	HMDB0036399	0.0381	0.22	1.00*	0.13
	3-(2H-1,3-benzodioxol-5-yl)-3-hydroxy-1-(4-methoxy-1-benzofuran-5-yl)propan-1-one	H	HMDB0129366	0.0347	0.22	1.00*	0.13
	3-[2,4-dihydroxy-3-(3-methylbut-2-en-1-yl)phenyl]-1-(3,4-dihydroxyphenyl)prop-2-en-1-one	H	HMDB0124802	0.0017	0.15	1.00*	0.13
	3-[3,4-dihydroxy-5-(3-methylbut-2-en-1-yl)phenyl]-1-(2,4-dihydroxyphenyl)prop-2-en-1-one	H	HMDB0135880	0.0017	0.15	1.00*	0.13
	3-[4-hydroxy-2-methoxy-3-(3-methylbut-2-en-1-yl)phenyl]-1-(3-hydroxyphenyl)propan-1-one	H	HMDB0124820	0.0381	0.22	1.00*	0.13
	3-[4-hydroxy-3-(3-methylbut-2-en-1-yl)phenyl]-1-(2,4,5-trihydroxyphenyl)prop-2-en-1-one	H	HMDB0135879	0.0017	0.15	1.00*	0.13
	3-O-(4-O-Methyl-a-D-glucopyranuronosyl)-L-arabinose	H	HMDB0039888	0.0288	0.28	1.00*	0.13
	3,4,5-trihydroxy-6-(2-methyl-3-oxo-1-phenylpropoxy)oxane-2-carboxylic acid	H	HMDB0133639	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-(2-oxo-4-phenylbutoxy)oxane-2-carboxylic acid	H	HMDB0133767	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-(3-oxo-1-phenylbutoxy)oxane-2-carboxylic acid	H	HMDB0133769	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-(3-phenyloxirane-2-carbonyloxy)oxane-2-carboxylic acid	H	HMDB0126548	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(2-methyl-3-phenyloxiran-2-yl)methoxy]oxane-2-carboxylic acid	H	HMDB0133653	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(2-methyl-3-phenylpropanoyl)oxy]oxane-2-carboxylic acid	H	HMDB0133675	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(2-oxo-3-phenylpropanoyl)oxy]oxane-2-carboxylic acid	H	HMDB0132501	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(2,3,5-trihydroxy-6-methyloxan-4-yl)oxy]oxane-2-carboxylic acid	H	HMDB0124752	0.0288	0.28	1.00*	0.13

Table continued (p. 18 / 20)

Unknown	Candidate	Pool	Database ID	$\triangle$Mass	MSTC	MSOC	MFTC
X-24838	3,4,5-trihydroxy-6-[(2E)-3-hydroxy-2-(phenylmethylidene)propoxy]oxane-2-carboxylic acid	H	HMDB0133645	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(4-methoxy-1-benzofuran-6-yl)oxy]oxane-2-carboxylic acid	H	HMDB0129410	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[1-(3-phenyloxiran-2-yl)ethoxy]oxane-2-carboxylic acid	H	HMDB0133742	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[2-hydroxy-3-(3-oxoprop-1-en-1-yl)phenoxy]oxane-2-carboxylic acid	H	HMDB0134051	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[2-hydroxy-6-(3-oxoprop-1-en-1-yl)phenoxy]oxane-2-carboxylic acid	H	HMDB0134052	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[2-methoxy-4-(prop-2-en-1-yl)phenoxy]oxane-2-carboxylic acid	H	HMDB0135244	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[3-(2-methyl-3-oxopropyl)phenoxy]oxane-2-carboxylic acid	H	HMDB0133641	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[3-(3-oxobutyl)phenoxy]oxane-2-carboxylic acid	H	HMDB0133771	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[4-(2-methyl-3-oxopropyl)phenoxy]oxane-2-carboxylic acid	H	HMDB0133643	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[4-(3-oxobutyl)phenoxy]oxane-2-carboxylic acid	H	HMDB0133773	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[4-hydroxy-2-(3-oxoprop-1-en-1-yl)phenoxy]oxane-2-carboxylic acid	H	HMDB0134057	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[4-hydroxy-3-(3-oxoprop-1-en-1-yl)phenoxy]oxane-2-carboxylic acid	H	HMDB0134056	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(2E)-3-(3-hydroxyphenyl)prop-2-enoyl]oxyoxane-2-carboxylic acid	H	HMDB0124974	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(2E)-3-(4-hydroxyphenyl)prop-2-enoyl]oxyoxane-2-carboxylic acid	H	HMDB0125165	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(3E)-1-hydroxy-4-phenylbut-3-en-2-yl]oxyoxane-2-carboxylic acid	H	HMDB0133733	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[(3E)-2-hydroxy-4-phenylbut-3-en-1-yl]oxyoxane-2-carboxylic acid	H	HMDB0133732	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[3-(2-hydroxyphenyl)prop-2-enoyl]oxyoxane-2-carboxylic acid	H	HMDB0134050	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[3-(3-hydroxyphenyl)prop-2-enoyl]oxyoxane-2-carboxylic acid	H	HMDB0127984	0.0500	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-[3-(4-hydroxyphenyl)prop-2-enoyl]oxyoxane-2-carboxylic acid	H	HMDB0128092	0.0500	0.27	1.00*	0.13

Table continued (p. 19 / 20)

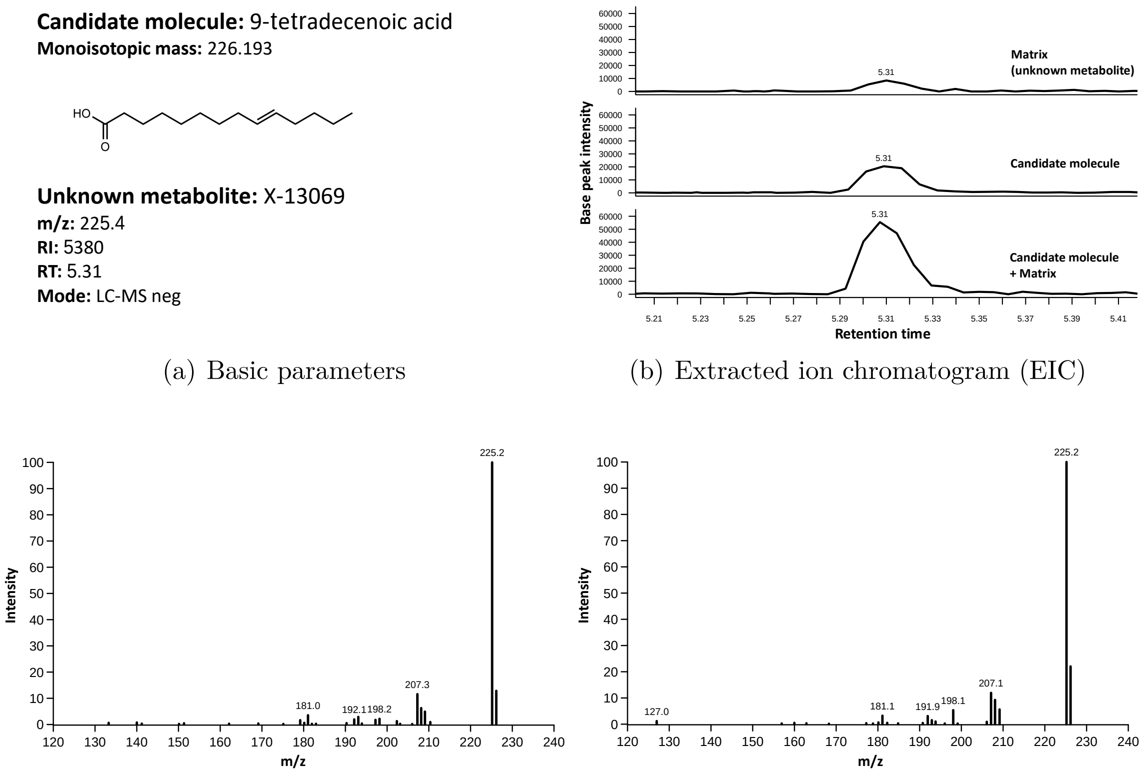
Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-24838	3,4,5-trihydroxy-6-[3-(4-methoxyphenyl)prop-2-en-1-yl]oxyoxane-2-carboxylic acid	H	HMDB0135652	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-2-[(1E)-3-hydroxy-2-methylprop-1-en-1-yl]phenoxyoxane-2-carboxylic acid	H	HMDB0133657	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-3-[(1E)-3-hydroxy-2-methylprop-1-en-1-yl]phenoxyoxane-2-carboxylic acid	H	HMDB0133647	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-3-[(1E)-3-hydroxybut-1-en-1-yl]phenoxyoxane-2-carboxylic acid	H	HMDB0133736	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-4-[(1E)-3-hydroxy-2-methylprop-1-en-1-yl]phenoxyoxane-2-carboxylic acid	H	HMDB0133650	0.0136	0.27	1.00*	0.13
	3,4,5-trihydroxy-6-4-[(1E)-3-hydroxybut-1-en-1-yl]phenoxyoxane-2-carboxylic acid	H	HMDB0133739	0.0136	0.27	1.00*	0.13
	3'-Ketolactose	H	HMDB0001030	0.0288	0.23	1.00*	0.13
	4-[1-hydroxy-3-(5-methoxy-2,2-dimethyl-2H-chromen-6-yl)propyl]phenol	H	HMDB0125851	0.0381	0.22	1.00*	0.13
	4-3-[4-hydroxy-3-(3-methylbut-2-en-1-yl)phenyl]oxirane-2-carbonylbenzene-1,3-diol	H	HMDB0135877	0.0017	0.22	1.00*	0.13
	4-Deacetylneosolaniol	H	HMDB0036158	0.0228	0.36	1.00*	0.13
	4-O-(4-O-Methyl-alpha-D-glucopyranuronosyl)-L-arabinose	H	HMDB0039725	0.0288	0.28	1.00*	0.13
	4-O-beta-D-Glucopyranuronosyl-L-fucose	H	HMDB0039889	0.0288	0.28	1.00*	0.13
	4',7-Dihydroxy-2'-methoxy-3'-prenylisoflavan	H	HMDB0034010	0.0381	0.18	1.00*	0.13
	5-Deoxykievitone	H	HMDB0034214	0.0017	0.18	1.00*	0.13
	5-Deoxymyricanone	H	HMDB0030799	0.0381	0.34	1.00*	0.13
	6-(2-benzyl-3-oxopropoxy)-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0133636	0.0136	0.27	1.00*	0.13
	6-[2-(2-carboxyeth-1-en-1-yl)phenoxy]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0134049	0.0500	0.27	1.00*	0.13
	6-[2-(3-formyloxiran-2-yl)phenoxy]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0134059	0.0500	0.27	1.00*	0.13
	6-[3-(2-carboxyeth-1-en-1-yl)phenoxy]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0127983	0.0500	0.27	1.00*	0.13
	6-[4-(2-carboxyeth-1-en-1-yl)phenoxy]-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0128091	0.0500	0.27	1.00*	0.13
	6-2-[(1E)-2-carboxyeth-1-en-1-yl]phenoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0141335	0.0500	0.27	1.00*	0.13
	6-3-[(1E)-2-carboxyeth-1-en-1-yl]phenoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0124973	0.0500	0.27	1.00*	0.13
	6-4-[(1E)-2-carboxyeth-1-en-1-yl]phenoxy-3,4,5-trihydroxyoxane-2-carboxylic acid	H	HMDB0125164	0.0500	0.27	1.00*	0.13

Table continued (p. 20 / 20)

Unknown	Candidate	Pool	Database ID	△Mass	MSTC	MSOC	MFTC
X-24838	7-(3,4-dihydroxyphenyl)-1-(3-methoxyphenyl)hept-1-ene-3,5-dione	H	HMDB0133497	0.0017	0.26	1.00*	0.13
	7-(4-hydroxy-3-methoxyphenyl)-1-(3-hydroxyphenyl)hept-1-ene-3,5-dione	H	HMDB0133502	0.0017	0.26	1.00*	0.13
	8-Acetyl-T2 tetrol	H	HMDB0036160	0.0228	0.36	1.00*	0.13
	Aesculin	H	HMDB0030820	0.0500	0.23	1.00*	0.13
	Cichoriin	H	HMDB0030821	0.0500	0.23	1.00*	0.13
	Citalopram N-oxide	H	HMDB0060654	0.0293	0.18	1.00*	0.13
	Desmethylxanthohumol	H	HMDB0030610	0.0017	0.15	1.00*	0.13
	Dihydro-O-methylsterigmatocystin	H	HMDB0030591	0.0347	0.18	1.00*	0.13
	Dolichin B	H	HMDB0029468	0.0017	0.22	1.00*	0.13
	Glyceollidin I	H	HMDB0034026	0.0017	0.20	1.00*	0.13
	Glyceollidin II	H	HMDB0037916	0.0017	0.20	1.00*	0.13
	Licocoumarone	H	HMDB0038755	0.0017	0.15	1.00*	0.13
	Linocinnamarin	H	HMDB0030678	0.0136	0.23	1.00*	0.13
	Lucanthone	H	HMDB0015607	0.0315	0.22	1.00*	0.13
	Morachalcone A	H	HMDB0040306	0.0017	0.15	1.00*	0.13
	Oenanthoside A	H	HMDB0035441	0.0136	0.23	1.00*	0.13
	Propiomazine	H	HMDB0014915	0.0315	0.22	1.00*	0.13
	Selinone	H	HMDB0037585	0.0017	0.22	1.00*	0.13
	Sulindac sulfide	H	HMDB0060614	0.0361	0.23	1.00*	0.13
	trans-isoeugenol-O-glucuronide	H	HMDB0060021	0.0136	0.27	1.00*	0.13
	Vulgaxanthin II	H	HMDB0029841	0.0387	0.26	1.00*	0.13
X-24860	Biotin	R	PCID 171548	0.0466	0.41	0.75	0.58*
	Carbidopa	H	HMDB0014336	0.0289	0.67*	0.92*	0.22
	gamma-Glutamylproline	H	HMDB0029157	0.0289	0.62	0.92*	0.22
	Glutamylproline	H	HMDB0028827	0.0289	0.62	0.92*	0.22
	Hydroxyprolyl-Hydroxyproline	H	HMDB0028864	0.0289	0.62	0.92*	0.22
	Hydroxyprolyl-Isoleucine	H	HMDB0028866	0.0075	0.70*	0.92*	0.22
	Hydroxyprolyl-Leucine	H	HMDB0028867	0.0075	0.70*	0.92*	0.22
	Isoleucyl-Hydroxyproline	H	HMDB0028908	0.0075	0.62	0.92*	0.22
	Isoleucyl-Isoleucine	H	HMDB0028910	0.0439	0.70*	0.92*	0.22
	Isoleucyl-Leucine	H	HMDB0028911	0.0439	0.70*	0.92*	0.22
	Leucyl-Hydroxyproline	H	HMDB0028930	0.0075	0.62	0.92*	0.22
	Leucyl-Isoleucine	H	HMDB0028932	0.0439	0.70*	0.92*	0.22
	Leucyl-Leucine	H	HMDB0028933	0.0439	0.70*	0.92*	0.22
X-24969	L-Cysteinylglycine disulfide	H	HMDB0000709	0.0069	0.61	1.00*	0.00
X-24983	D-argininium(1+)	R	PCID 71070	0.0320	0.57	0.80	0.60*

Table B.11: Automatically selected candidate molecules
676 candidate molecules (67 within the Recon-pool (R) and further 609 within the HMDB-pool (H)) have been automatically selected for 139 unknown metabolites based structural similarity to neighboring known metabolites of our network model. Database identifier refer to HMDB or PubChem. The last three columns contain the maximum values of the fingerprint Tanimoto (MFTC), structure Tanimoto (MSTC), and structure overlap coefficients (MSOC). Asterisks point on values passing the threshold. A MFTC of 0.00 indicates that the fingerprint of the related candidate was not available. A comprehensive version of this table including all evidences is provided in Supplemental File B.8.

Experimental verification of candidate molecules



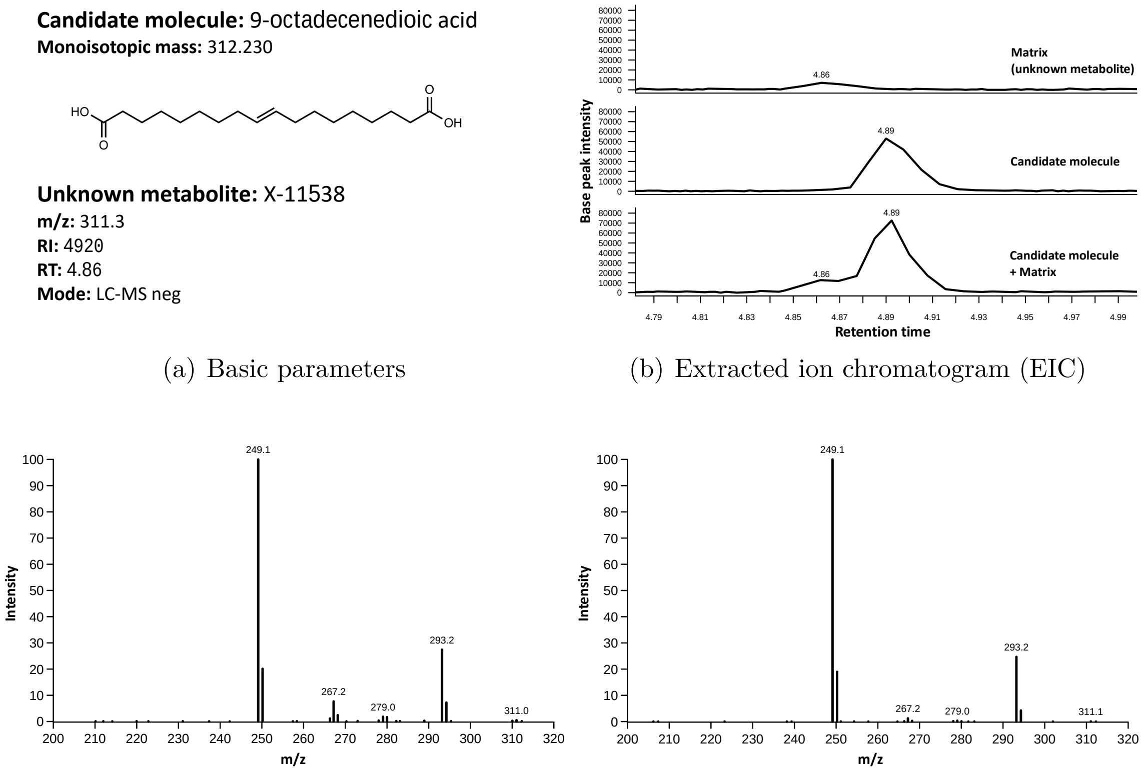
(a) Basic parameters

(b) Extracted ion chromatogram (EIC)

(c) MS^2 fragmentation spectrum: candidate

(d) MS^2 fragmentation spectrum: X-13069

Figure B.6: Analytical comparison of 9-tetradecenoic acid and X-13069
The EIC of each aliquot shows the same retention time: candidate molecule, reference matrix (containing the unknown metabolite), and candidate molecule + reference matrix. Both MS^2 fragmentation spectra of the candidate molecule and the unknown metabolite show the same fragments with equal relative intensities leading to a verification of the candidate molecule. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].



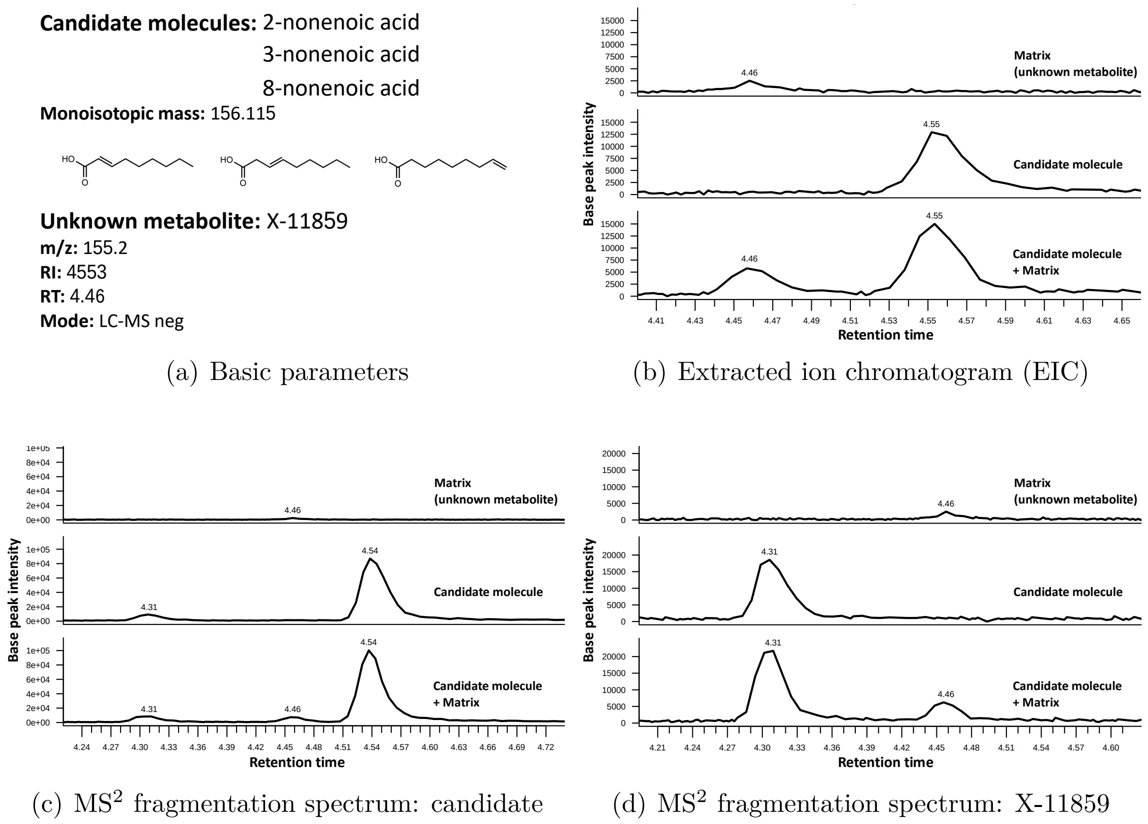
(a) Basic parameters

(b) Extracted ion chromatogram (EIC)

(c) MS^2 fragmentation spectrum: candidate

(d) MS^2 fragmentation spectrum: X-11538

Figure B.7: Analytical comparison of 9-octadecenedioic acid and X-11538
The EIC of 9-octadecenedioic acid and X-11538 show slightly shifted retention times (4.89 and 4.86 min). Both MS^2 fragmentation spectra are similar, but not identical: two main peaks at 249.1 and 293.2 show equal relative intensities, but two smaller peaks at 267.2 and 279.1 are much larger for the plasma compared to the candidate aliquot. In summary this candidate must be falsified. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].



(a) Basic parameters

(b) Extracted ion chromatogram (EIC)

(c) MS^2 fragmentation spectrum: candidate

(d) MS^2 fragmentation spectrum: X-11859

Figure B.8: Analytical comparison of 2-, 3-, and 8-nonenoic acid and X-11859
Candidates 2-nonenoic acid, 3-noneoic acid, or 8-noneoic acid and the unknown metabolite X-11859 have different retention time peaks in their EICs. Partially, there common mass-to-charge ratios in the MS^2 fragmentation spectra (not shown), but their relative intensities are different. In summary these candidates must be falsified. I previously published this figure in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Unknown	Candidate	m/z	RT$_U$	RT$_C$	MS^2_U signals	MS^2_C signals	Match
X-13891	2-dodecenedioic acid	227.1	2.77	2.77	165.1, 183.0, 209.1	165.0, 183.0, 209.1	yes
X-13069	9-tetradecenoic acid	225.4	5.31	5.31	225.2, 207.1	225.2, 207.3	yes
X-11538	9-octadecenedioic acid	311.3	4.86	4.89	249.1, 293.2, 267.2, 279.0	249.1, 293.2, 267.2, 279.0	no
X-11859	2-nonenoic acid	155.2	4.46	4.55	110.8, 82.0, 155.2, 123.1, 136.9	155.0, 111.1	no
X-11859	3-nonenoic acid	155.2	4.46	4.54	110.8, 82.0, 155.2,	111.1, 155.0, 137.1	no
				4.31	123.1, 136.9	155.0, 137.0	no
X-11859	8 nonenoic acid	155.2	4.46	4.31	110.8, 82.0, 155.2, 123.1, 136.9	155.0, 137.0	no

Table B.12: Evaluation of manually selected candidate molecules
RT and main MS^2 fragments of LC-MS measurements are juxtaposed for manually selected candidate molecules (C) and respective unknown metabolites (U), ordered by intensity descending. Both characteristics match for two candidate molecules (verification) and differs in four cases (falsification). I previously published this table in Quell et al. 2017; reprinted by permission from the Journal of Chromatography B [150].

Supplemental files

Supplemental files of Chapter 4 are provided as online-supplements and on the attached CD (cf. Index of external material on page 239).

Part_II_UnknownAnnotation/Supplemental_File_B.1_Network_Model.rar

File B.1: Toolbox with R scripts and example data
The zip file contains implementations for all modules of the metabolite characterization approach in R along with input data, sample outputs, and a instruction. I published parts of this file previously in Quell et al. 2017 [150]. This version is extended by major parts, such as by the automated candidate selection.

Part_II_UnknownAnnotation/
Supplemental_File_B.2_Network_SUMMIT.graphml

File B.2: Graphml file of a representation of the network model based on SUMMIT
The network model embeds 254 measured unknown metabolites into their biochemical and functional context. The model was constructed based on 758 measured metabolites (439 known, 319 unknown), 2 626 metabolites of the public database Recon and 1 782 genes. For clarity, elements of Recon are incorporated only if they are either directly or through a gene connected to a measured metabolite. Not-connected metabolites and genes are not shown. Edge and node colors are equal to the network representation in Figure 4.1. I previously published this file in Quell et al. 2017; republished by permission from the Journal of Chromatography B [150].

Part_II_UnknownAnnotation/
Supplemental_File_B.3_unknownAnnotations_SUMMIT.xlsx

File B.3: Automatically generated annotations of the SUMMIT data set
The automatic workflow annotated 180 unknown metabolites with predicted pathways and selected reactions for 109 unknown metabolites based on 2 279 blood samples of SUMMIT. Moreover, the file contains information regarding the measurement of unknown metabolites and neighboring metabolites and genes within the network model. I previously published this file in Quell et al. 2017; republished by permission from the Journal of Chromatography B [150].

Part_II_UnknownAnnotation/
Supplemental_File_B.4_Network_GCKD.graphml

File B.4: Graphml file of a representation of the network model based on GCKD
The network model embeds 482 measured unknown metabolites into their biochemical and functional context. The model was constructed based on 1 456 measured metabolites (796 known, 660 unknown), 706 metabolites of the public database Recon and 435 genes. For clarity, elements of Recon are incorporated only if they are either directly or through a gene connected to a measured metabolite. Not-connected metabolites and genes are not shown. Edge and node colors are equal to the network representation in Figure 4.1.

Part_II_UnknownAnnotation/
Supplemental_File_B.5_unknownAnnotations_GCKD.xlsx

File B.5: Automatically generated annotations of the GCKD data set
The automatic workflow annotated 273 unknown metabolites with predicted pathways and selected reactions for 297 unknown metabolites based on 1 209 blood samples of GCKD. Moreover, the file contains information regarding the measurement of unknown metabolites and neighboring metabolites and genes within the network model.

Part_II_UnknownAnnotation/
Supplemental_File_B.6_Network_SUMMIT_GCKD.graphml

File B.6: Graphml file of the network model based on merged SUMMIT/GCKD
The network model embeds 677 measured unknown metabolites into their biochemical and functional context. The model was constructed based on 1 824 measured metabolites (984 known, 840 unknown), 865 metabolites of the public database Recon and 529 genes. For clarity, elements of Recon are incorporated only if they are either directly or through a gene connected to a measured metabolite. Not-connected metabolites and genes are not shown. Edge and node colors are equal to the network representation in Figure 4.1.

Part_II_UnknownAnnotation/
Supplemental_File_B.7_unknownAnnotations_SUMMIT_GCKD.xlsx

File B.7: Automatically generated annotations of merged SUMMIT/GCKD
The automatic workflow annotated 435 unknown metabolites with predicted pathways and selected reactions for 384 unknown metabolites based on the merged sets of SUMMIT and GCKD. Moreover, the file contains information regarding the measurement of unknown metabolites and neighboring metabolites and genes within the network model.

Part_II_UnknownAnnotation/

Supplemental_File_B.8_autoSelectedCandidates.xlsx

File B.8: Evidences for automatically selected candidate molecules
676 candidate molecules (67 within the Recon-pool (R) and further 609 within the HMDB-pool (H)) have been automatically selected for 139 unknown metabolites based structural similarity to neighboring known metabolites of our network model. Fingerprint Tanimoto, structure Tanimoto, and structure overlap coefficients are specified for each known neighbor of unknown metabolites to the respective database compound. A fingerprint Tanimoto coefficient of 0.00 indicates that the fingerprint of the related candidate was not available. the Evidences column summarizes information that supports the candidate molecule. An overview on automatically selected candidate molecules is provided in Table B.11.

Index of external materials

Supplemental materials are provided on the attached CD and as online-supplement (stable URL: `https://DissertationQuell.page.link/Supplements`).

Description	**Path**
Supplemental material of Chapter 3	Part_I_LipidCompositions/ Supplemental_
Suppl. Table A.1: Qualitative composition of PCs measured by Absolute*IDQ*$^{\text{TM}}$	..Table_A.1_QualitativeComp.xlsx
Suppl. Table A.2: Qualitative composition of measured PCs and ranges of metabolite levels per platform	..Table_A.2_Ranges.xlsx
Suppl. Table A.3: Stability of factors f describing PC compositions	..Table_A.3_Stability.xlsx
Suppl. Table A.4: Replication of factors f and imputation in QMDiab	..Table_A.4_Replication.xlsx
Suppl. Table A.5: Explained variation of PCs measured on the lipid species level	..Table_A.5_LinearModel.xlsx
Suppl. Figure A.1: Details of PC compositions	..Figure_A.1_CompDetails.pdf
Suppl. Figure A.2: Factors f in HuMet and QMDiab of PC compositions	..Figure_A.2_Replication_factor.pdf
Suppl. Figure A.3: Distributions of measured and imputed PC concentrations	..Figure_A.3_Replication_conz.pdf

Description	Path
Supplemental material of Chapter 4	Part_II_UnknownAnnotation/Supplemental_
Suppl. File B.1: Network model with R scripts and example data	..File_B.1_Network_Model.rar
Suppl. File B.2: Graphml file of a representation of the network model based on SUMMIT	..File_B.2_Network_SUMMIT.graphml
Suppl. File B.3: Automatically generated annotations of the SUMMIT data set	..File_B.3_unknownAnnotations_SUMMIT.xlsx
Suppl. File B.4: Graphml file of a representation of the network model based on GCKD	..File_B.4_Network_GCKD.graphml
Suppl. File B.5: Automatically generated annotations of the GCKD data set	..File_B.5_unknownAnnotations_GCKD.xlsx
Suppl. File B.6: Graphml file of the network model based on merged SUMMIT/GCKD	..File_B.6_Network_ SUMMIT_GCKD.graphml
Suppl. File B.7: Automatically generated annotations of merged SUMMIT/GCKD	..File_B.7_unknownAnnotations_ SUMMIT_GCKD.xlsx
Suppl. File B.8: Evidences for automatically selected candidate molecules	..File_B.8_autoSelectedCandidates.xlsx

List of figures

List of tables

List of abbreviations

ANOVA	analysis of variance
BMI	body mass index
CN	correlation network
CV	coefficient of variation
DMS	differential ion mobility spectrometry
EIC	extracted ion chromatogram
ESI	electrospray ionization
FIA	flow injection analysis
FN	false negative
FP	false positive
GC	gas chromatography
GCKD	German chronic kidney disease
GGM	Gaussian graphical model
HILIC	hydrophilic interaction liquid chromatography
HMDB	Human Metabolome Database
HuMet	Human Metabolome
InChI	International Chemical Identifier
KEGG	Kyoto Encyclopedia of Genes and Genomes
KORA	Cooperative Health Research in the Augsburg Region
LC	liquid chromatography
lysoPC	lysophosphatidylcholine
MALDI	matrix-assisted laser desorption/ ionization
MCS	maximum common substructure
MFTC	maximum fingerprint Tanimoto coefficient
mGWAS	metabolite-based genome-wide association study
MRM	multiple reaction monitoring
MS	mass spectrometry

MS/MS	tandem mass spectrometry
MSOC	maximum MCS overlap coefficient
MSTC	maximum MCS Tanimoto coefficient
MWAS	metabolome-wide association study
NAKO	national cohort
NMR	nuclear magnetic resonance
OGTT	oral glucose tolerance test
OLTT	oral lipid tolerance test
PAT	physical activity test
PC	phosphatidylcholine
ppm	parts per million
QC	quality control
QMDiab	Qatar Metabolomics Study on Diabetes
RI	retention index
RP	reversed phase
RT	retention time
SD	standard deviation
SDF	structure data file
SLD	standard liquid diet
SM	sphingomyelin
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry System
SNP	single nucleotide polymorphism
SUMMIT	surrogate markers for micro- and macro-vascular hard endpoints for innovative Diabetes tools
TIC	total ion chromatogram
TN	true negative
TOF	time of flight
TP	true positive
XML	extensible markup language

Teaching activity

Taking care of training students was one of my special concerns during my early stage career. I organized, prepared and/ or performed the following lectures, seminars and practical trainings that were open to bachelor or master students in Bioinformatics, Informatics, Biology, or Molecular Biotechnology of the Technical University of Munich and (co-) supervised the following theses. Furthermore, I absolved didactic trainings to receive the certificate 'Hochschullehre der Bayerischen Universitäten – Vertiefungsstufe'.

Summer term 2018

- Exercise course 'Einführung in die Bioinformatik II'
- Guest lecture in 'Einführung in die Bioinformatik II', topic: protein structure, June 2018

Winter term 2017/ 18

- Exercise course 'Einführung in die Bioinformatik I'

Summer term 2017

- Exercise course 'Einführung in die Bioinformatik II'
- Guest lecture in 'Einführung in die Bioinformatik II', topic: protein structures and -prediction, May 2017
- Supervision of the thesis 'Gender differences in the effect of medications on human blood metabolomes', Luise Schuller

Winter term 2016/ 17

- Exercise course ‘Einführung in die Bioinformatik I’
- Practical training ‘Genomorientierte Bioinformatik’

Summer term 2016

- Exercise course ‘Einführung in die Bioinformatik II’
- Practical training ‘Masterpraktikum Bioinformatik’
- Practical training ‘Angewandte Bioinformatik’
- Guest lecture in ‘Einführung in die Bioinformatik II’, topic: protein structures and -prediction, Juni 2016

Winter term 2015/ 16

- Exercise course ‘Einführung in die Bioinformatik I’
- Practical training ‘Genomorientierte Bioinformatik’

Summer term 2015

- Supervision of a student’s group for the practical training ‘Masterpraktikum Bioinformatik’

Summer term 2014

- Supervision of the thesis ‘Prioritizing pathways and genes linked to rare mitochondriopathies based on metabolic extremes in individual patients’, Rene Schoeffel

Winter term 2013/ 14

- Exercise course ‘Einführung in die Bioinformatik I’
- Supervision of a student’s group for the practical training ‘Genomorientierte Bioinformatik’