Automated pathway and reaction prediction facilitates in silico identification of unknown metabolites in human cohort studies

Automated pathway and reaction prediction facilitates *in silico* identification of unknown metabolites in human cohort studies

Jan D. Quell[a,b], Werner Römisch-Margl[b], Marco Colombo[c], Jan Krumsiek[d], Anne M. Evans[e], Robert Mohney[e], Veikko Salomaa[f], Ulf de Faire[g], Leif C. Groop[h], Felix Agakov[i], Helen C. Looker[j], Paul McKeigue[c], Helen M. Colhoun[j], Gabi Kastenmüller[b,*]

Affiliations:
[a] Genome-oriented Bioinformatics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany
[b] Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[c] Centre for Population Health Sciences, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, UK
[d] Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[e] Metabolon Inc., Durham, North Carolina, USA
[f] Department of Health, National Institute for Health and Welfare, Finland
[g] Unit of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[h] Department of Clinical Sciences - Diabetes and Endocrinology, Lund University, Malmö, Sweden
[i] Pharmatics Limited, Edinburgh, UK
[j] Population Health Sciences, University of Dundee, Dundee, UK
[*] corresponding author

E-Mail:
g.kastenmueller@helmholtz-muenchen.de

Abstract

Identification of metabolites in non-targeted metabolomics continues to be a bottleneck in metabolomics studies in large human cohorts. Unidentified metabolites frequently emerge in the results of association studies linking metabolite levels to, for example, clinical phenotypes. For further analyses these unknown metabolites must be identified. Current approaches utilize chemical information, such as spectral details and fragmentation characteristics to determine components of unknown metabolites. Here, we propose a systems biology model exploiting the internal correlation structure of metabolite levels in combination with existing biochemical and genetic information to characterize properties of unknown molecules.

Levels of 758 metabolites (439 known, 319 unknown) in human blood samples of 2279 subjects were

measured using a non-targeted metabolomics platform (LC-MS and GC-MS). We reconstructed the structure of biochemical pathways that are imprinted in these metabolomics data by building an empirical network model based on 1040 significant partial correlations between metabolites. We further added associations of these metabolites to 134 genes from genome-wide association studies as well as reactions and functional relations to genes from the public database Recon 2 to the network model. From the local neighborhood in the network, we were able to predict the pathway annotation of 180 unknown metabolites. Furthermore, we classified 100 pairs of known and unknown and 45 pairs of unknown metabolites to 21 types of reactions based on their mass differences. As a proof of concept, we then looked further into the special case of predicted dehydrogenation reactions leading us to the selection of 39 candidate molecules for 5 unknown metabolites. Finally, we could verify 2 of those candidates by applying LC-MS analyses of commercially available candidate substances. The formerly unknown metabolites X-13891 and X-13069 were shown to be 2-dodecendioic acid and 9-tetradecenoic acid, respectively.

Our data driven approach based on measured metabolite levels and genetic associations as well as information from public resources can be used alone or together with methods utilizing spectral patterns as a complementary, automated and powerful method to characterize unknown metabolites.

Keywords

Metabolite identification
Non-targeted metabolomics
Biochemical pathway prediction
Reaction prediction
Metabolic network reconstruction

1 Introduction

Non-targeted metabolomics based on liquid chromatography coupled to mass spectrometry (LC-MS) has emerged as an established technology to simultaneously measure the levels of a wide range of low weight molecules (metabolites) in biofluids and tissues [1]. While the non-targeted approach allows the discovery

of unexpected metabolic links in many fields of biomedical research [2], a significant fraction of the obtained analytical signals cannot be assigned to a chemical structure though they are stably measured in thousands of samples [3]. Two years ago, the Metabolite Identification Task Group of the Metabolomics Society accentuated the community consensus that identification of these so-called unknown metabolites measured by several non-targeted mass spectrometry techniques in a larger scale is one of the most significant current challenges in metabolomics [4,5].

Traditional identification of unknown metabolites in wet laboratories is very expensive and time consuming. Consequently, the attempt of identifying metabolites *in silico* was started as research niche a couple of years ago, and is more and more becoming a hot topic in metabolomics [6]. Various current *in silico* approaches focus on fragmentation spectra of unknown metabolites. As an example, Allen et al. published a probabilistic model, called Competitive Fragmentation Modeling (CFM) that uses fragmentation graphs and machine learning techniques to reproduce the unknown fragmentation based on known spectra of known chemical structures or to predict and rank possible structures based on a mass spectrum [7]. Recently, Ruttkies et al. used *in silico* fragmentation (MetFrag) and calculation of the retention time to evaluate candidates for unknown metabolites [8]. Grapov et al. proposed a graph-based tool, called MetaMapR, that integrates a similarity measure based on mass spectra with database information, such as enzymatic transformations and metabolite structural similarity to achieve richly connected metabolic networks incorporating unknown metabolites [9].

Following a different idea without using spectral features, we previously suggested a systems biology method for the identification of unknown metabolites that is primarily based on the (partial) correlation between measured concentrations of metabolites and their genetic associations determined in metabolomics data from large cohorts [10]. We demonstrated that the network of metabolite pairs with significant partial correlation, the so-called Gaussian graphical models (GGMs), reconstruct biochemical pathways from metabolomics data [11]. By combining GGMs with metabolite-gene associations from genome-wide association studies with metabolites as quantitative traits (mGWAS) in a network, we were able to retrieve biochemical, functional information for unknown metabolites through manual inspection of

the resulting network. For further manual look-up, we provided Gene Ontology terms as well as known biochemical reactions from metabolic databases in annotation tables for genes and measured metabolites. This further facilitated the characterization of unknown metabolites [10].

Here, we extend this idea by (i) directly augmenting the GGM- and GWAS-based network with metabolite-metabolite and metabolite-gene links from prior knowledge on biochemical reactions as stored in public databases such as Recon 2 [12] to make this knowledge accessible for systematic and automated mining, and by (ii) providing systematic and automated downstream analysis of the final integrated network replacing its manual inspection. To this end, we predict the biochemical pathways of the unknown metabolites in the network based on their neighbors with known chemical identity. In addition, we use mass differences between the unknown and neighboring known metabolites to predict enzymatic reactions of the unknown metabolites based on its measured mass to charge ratio (m/z) as previously proposed by Breitling et al. [13].

To demonstrate its applicability for metabolite identification, we apply our approach to a non-targeted metabolomics dataset (439 known, 319 unknown metabolites) from the blood samples of 2279 subjects that were analyzed in the course of the project “Surrogate markers for Micro- and Macro-vascular hard endpoints for Innovative diabetes Tools” (SUMMIT) using a commercial LC-MS-based metabolomics platform (Metabolon Inc., USA). For a selected group of predicted metabolites, for which the pure compounds were commercially available, we tested our predictions experimentally.

2 Materials and Methods

The procedure of our automated metabolite characterization approach consists of modules for GGM generation, for integration of public data, and for pathway and reaction prediction. Figure S1 shows an overview of the complete workflow. We demonstrated the applicability of our method using metabolomics data that was produced in the course of the SUMMIT project by non-targeted LC-MS analysis. Implementations of all modules in R are provided in Supplementary File S1 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 experimental validation.

2.1 Study cohorts and metabolomics data

Serum samples of n=2279 patients with type 2 diabetes (T2D) from seven population studies, FINRISK1997 (n=242), FINRISK2002 (n=92), FINRISK2007 (n=28) [14], Go-DARTS (n=1200) [15], IMPROVE (n=44) [16], 60-years-olds (n=20) [17] and SDR (n=653) [18], all participating in the SUMMIT project, were analyzed using the non-targeted metabolomics platform of Metabolon Inc. (Durham, USA). 1147 of the T2D patients were also diagnosed with cardiovascular disease (CVD) while 1132 did not suffer from CVD. Besides the T2D and CVD disease state of the patients, further clinical information such as age, sex, duration of type 2 diabetes, height, body mass index (BMI), triglyceride, HDL, LDL, DBP, SBP, smoking status, hemoglobin A1c, baseline estimated glomerular filtration rate, insulin status and medication information such as ACE inhibitors, angiotensin receptor blockers (ARB), calcium channel blockers (CCB), diuretics, lipid rx, blood pressure lowering drugs, beta blockers, alpha blockers and aspirin was available.

The non-targeted metabolomics platform comprises LC-MS (in positive and negative mode) as well as MS coupled to gas chromatography (GC) and has been described in detail previously [19,20]. 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 (2.1mm × 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 electrospray ionization source of the mass spectrometer. Full MS scans were recorded

from 80 to 1000 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 gas chromatography 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 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.

In total, the levels of 758 metabolites were determined for the 2279 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 the Kyoto Encyclopedia of Genes and Genomes (KEGG) [21].

2.2 Public data sources

For integrating metabolite-metabolite and metabolite-gene links based on known biochemical reactions into our model, we used Recon 2 [12], 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 [21] or HumanCyc [22]. Furthermore, we used a published Gaussian graphical model (GGM) based on the population cohort KORA F4 (n=1768) [10], and metabolite-gene associations of a published metabolomics GWAS based on KORA F4 and TwinsUK (n=6056) [23].

2.3 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 [24].

GGM generation: Based on the metabolomics data of the 2279 subjects of our cohorts, we created a GGM as backbone of our network model, since they are known to reconstruct biochemical pathways from measured metabolite levels [11]. 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 [25] package 'mice' [26] 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' [27] 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 incorporating 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 into one model to end up with 1040 connections between 637 known and unknown metabolites. Then we added 134 metabolite-associated genes of a published GWAS [23]. Finally, we attached knowledge-based biochemical information extracted from Recon 2 [12] to the network model. To this end, we first added metabolites from Recon 2 if they were functionally related to at least one of the 134 genes through a reaction listed in Recon 2. 343 Recon metabolites, of which 37 were mapped to measured metabolites and thus were already part of the GGM- and GWAS-based network, showed functional links to 57 genes, which were added to the network.

Secondly, Recon 2 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 2 reaction. In total, the resulting (final) network includes, 1152 Recon 2 reactions connecting 591 metabolites. In total, the resulting (final) network incorporates Please note that by integrating only main metabolites, which are annotated as 'baseReactants' and 'baseProducts' in Recon 2, 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 reaction is extracted from other resources that do not include such annotations [28–30]. Please also note that, for our purposes, we ignored the compartment annotations provided with metabolite species in Recon 2 reactions, i.e., each Recon metabolite mapped or added to the GGM- and GWAS-based network is represented by a single node and two metabolites are connected if they are linked through a Recon 2 reaction irrespective of the compartment in which the reaction takes place. Reactions that are classified as 'Transport' or 'Exchange' in Recon2 are omitted when integrating Recon reactions into the network. The data integration process is schematically visualized in Figure S2.

2.4 Prediction of super and sub pathways of unknown metabolites

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 unknown metabolites' biochemical context. 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 2b). To define the neighborhood for metabolites, we consider metabolites as neighbors, if they are connected by a GGM edge, share a common GWAS 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 2a).

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$\}$ (Formula (1)).

$$P_B(p) = \frac{\sum \text{\# neighbors of metabolites with super pathway } p}{2 \cdot \text{\# neighboring metabolite pairs}} \quad \textbf{(1)}$$

For each super pathway p, we calculated the conditional probability $P_N(p_i|p_j)$ based on all known metabolites i with super pathway p_i given metabolites j with super pathway p_j are neighbors of i (Formula (2).

$$P_N(p_i|p_j) = \frac{P(p_i \cap p_j)}{P_B(p_j)} = \frac{P(p_i \text{ and } p_j \text{ are neighbors})}{\text{background probability of } p_j} \quad \textbf{(2)}$$

During the prediction phase we resolve the conditional probability $P_N(p_i|p_1 \cap \ldots \cap p_n)$ for each super pathway p_i of all unknown metabolites i given the pathways $p_1, \ldots, p_n$ of n neighboring known metabolites

(Formula (3). The first transformation follows the Bayes' theorem. For the second transformation we assumed independence of the neighbors $1, \dots, n$. 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.

$$P_N(p_i|p_1 \cap \dots \cap p_n) \Leftrightarrow \frac{P_N(p_1 \cap \dots \cap p_n|p_i) \cdot P_B(p_i)}{P_B(p_1 \cap \dots \cap p_n)} \sim\propto \frac{P_N(p_1|p_i) \cdot \dots \cdot P_N(p_n|p_i) \cdot P_B(p_i)}{P_B(p_1) \cdot \dots \cdot P_B(p_n)} \quad \textbf{(3)}$$

For each unknown metabolite i, the predicted super pathway p_i with the highest probability $max(P_N(p_i|p_1 \cap \dots \cap p_n))$ 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 $\geq 97.5\% \Longrightarrow z \geq 207.0$), (b) high confidence (correct predictions $\geq 95\% \Longrightarrow z \geq 78.0$), (c) medium confidence (correct predictions $\geq 90\% \Longrightarrow z \geq 7.1$), (d) low confidence (correct predictions $\geq 85\% \Longrightarrow z \geq 2.7$) and (e) very low confidence (correct predictions $< 85\% \Longrightarrow z < 2.7$). 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 (Table S1).

2.5 Prediction of reactions that connect known and unknown metabolites

Knowledge about the reaction connecting a known and 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 2. 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 GWAS or via a known reaction according to Recon 2 (Figure 2c). 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 [11]. 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 $\Delta m_{pair} = \Delta m_{expected} \pm e$ [13]. Here, we set $e = 0.3$ to compensate the unit mass resolution of theMS platform, on which the presented data was collected. e can be adapted for metabolomics platforms with higher mass resolution to yield more specific reaction types. The predicted reaction can be applied to the known metabolite to manually select concrete candidate molecules for the connected unknown metabolite.

We evaluated our approach based on pairs of known metabolites (Table 2).

2.6 Experimental validation of candidate molecules

For a selected set of unknown metabolites, we sought experimental confirmation of our predictions. Here, we focused on unknown metabolites for which the most frequently observed reaction type, dehydrogenation reactions, were predicted. To select the most promising candidates for experimental validation, we additionally filtered the list using differences in retention index as a second criterion. The distribution of differences in retention indices 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 S3), we used the mean difference in retention time of the correct predictions plus/ minus their variances ($\Delta ri < 355.9$) as threshold for selecting 26 of the most promising candidates of unknown metabolites for experimental validation.

To verify or falsify these candidates of unknown metabolites we purchased all corresponding molecules that were commercially available as pure substances: 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-dodecendioic acid (VWR International GmbH, Darmstadt, Germany: CAYM88820, 228.14 g/mol).

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) coupled to a Waters Acquity UPLC system (Waters GmbH, Eschborn, Germany) at the Helmholtz Zentrum München. After sample injection the column (2.1mm × 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 retention time of peaks in the extracted ion chromatogram (EIC) for the three measurements per metabolite. Especially the mixture of the pure substance and reference plasma should show just one peak, because two separate peaks would indicate that both substances are not identical. Finally, we checked if the fragment spectra of the pure substances and of the respective unknown metabolite in the matrix sample consisted of the same fragments with equal relative intensities.

3 Results

Here, we propose a systems biology method for identification of unknown metabolites that is based on the investigation of the unknown metabolite's biochemical and functional neighborhood in a metabolic network that is reconstructed from the metabolomics data using Gaussian Graphical Models (GGM) (see Materials and Methods). We further extend the network by genetic associations and prior biochemical knowledge from public databases and use it as a basis for automatically predicting the biochemical pathways and, in various cases, the reaction by which the unknown metabolite is produced from a known metabolite. This procedure yields concrete molecules as candidates for unknown metabolites that can then be tested experimentally. We applied our approach to unknown metabolites in non-targeted metabolomics data from blood of 2279 subjects. As a proof of principle, we tested selected predicted candidates on the LC-MS metabolomics platform.

3.1 Data integration and construction of the network model

The network model is the core part of our approach and the basis of analytical and predictive methods. It connects unknown metabolites to complementary functional information from heterogeneous data resources 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 (see Methods). 637 (388 known, 249 unknown) of the 758 measured metabolites are connected by 1040 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 published metabolomics GWAS 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 GWAS 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 2. 591 of those metabolites are connected through 1152 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 2. In the final model, 181 unknown metabolites are connected by a GGM edge (171) or via a gene (35) to a known metabolite. A graphml file of the network model is prepared in Supplementary File S2.

The overall network (Figure 1), which shows a scale free topology, embeds unknown metabolites into their biochemical and functional context (Figure 1, zoom in) by connecting them to known metabolites via direct edges (GGM, blue) or indirect edges (GWAS, green; Recon 2, red). Thereby metabolites are connected neglecting the compartmentalization of biochemical processes. In contrast to networks aiming at a realistic reconstruction of complete human metabolism for modelling 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.

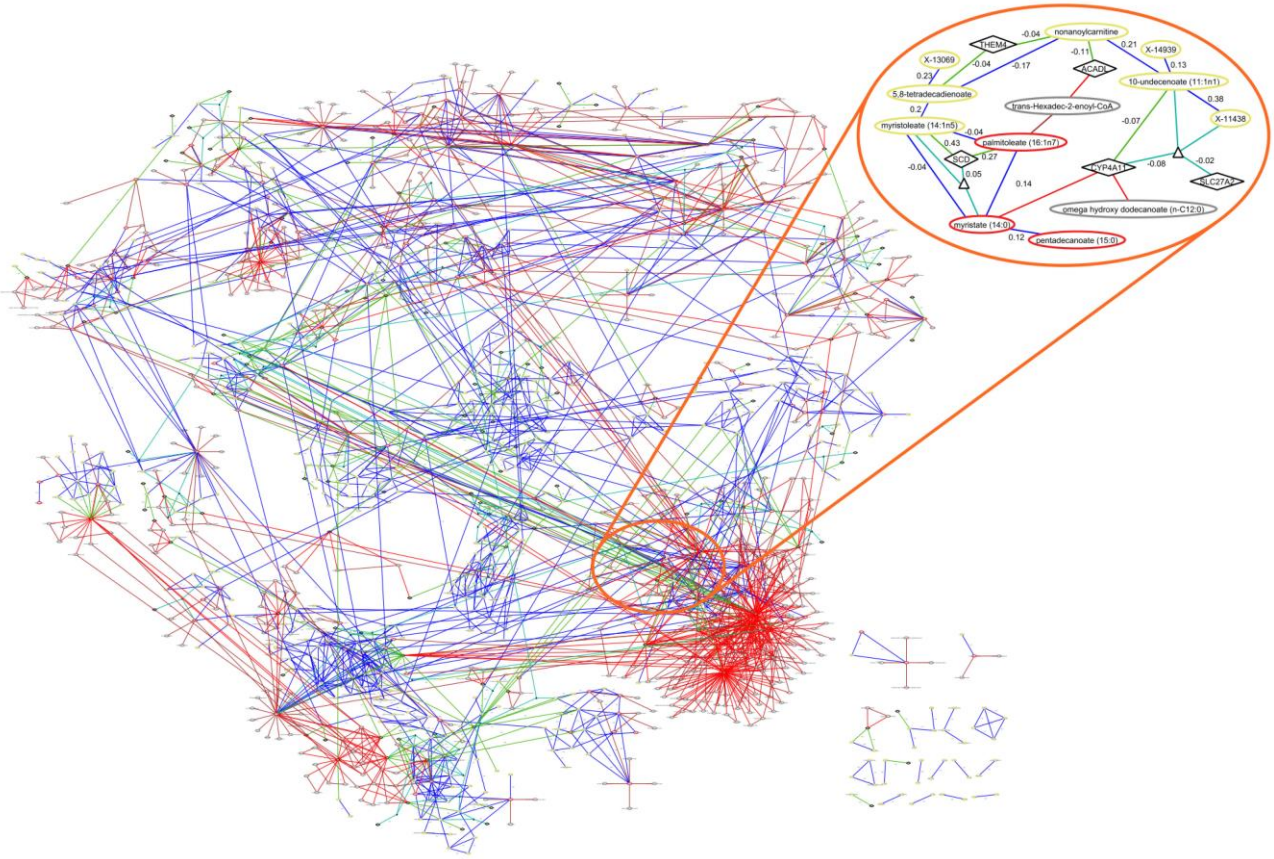


Figure 1: Graphical representation of the network model
The final network model embeds 254 measured unknown metabolites into their biochemical and functional context. The model was

constructed based on 758 measured metabolites (known: 439, unknown: 319), 2626 metabolites of the public database Recon 2 and 1782 genes from Recon 2 and published metabolomics GWAS. Elements of Recon 2 were included only if they were either directly or through a gene connected to a measured metabolite. Not-connected metabolites and genes are not shown. Edge colors indicate the type of connection as follows: blue: GGM, green: GWAS, red: functional relation (Recon), brown: reaction (Recon). GGM edges are labeled with the mass difference of metabolites and the beta is shown for GWAS 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).

3.2 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 2a and b). To 150 metabolites a clear sub pathway was assigned. For 28 metabolites two (mostly similar) sub pathways were suggested. Table 1 summarizes the proportion of predicted super and sub pathways per confidence class.

Confidence [%]	Confidence	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 1: Proportion of unknown metabolites with predicted pathways
The five confidence levels that are provided for the pathway predictions were determined based on thresholds for known metabolites (see Materials and Methods).

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 S1 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 pathways for unknown metabolites is provided in Table S2. Our approach is able to classify a large amount of unknown metabolites automatically.

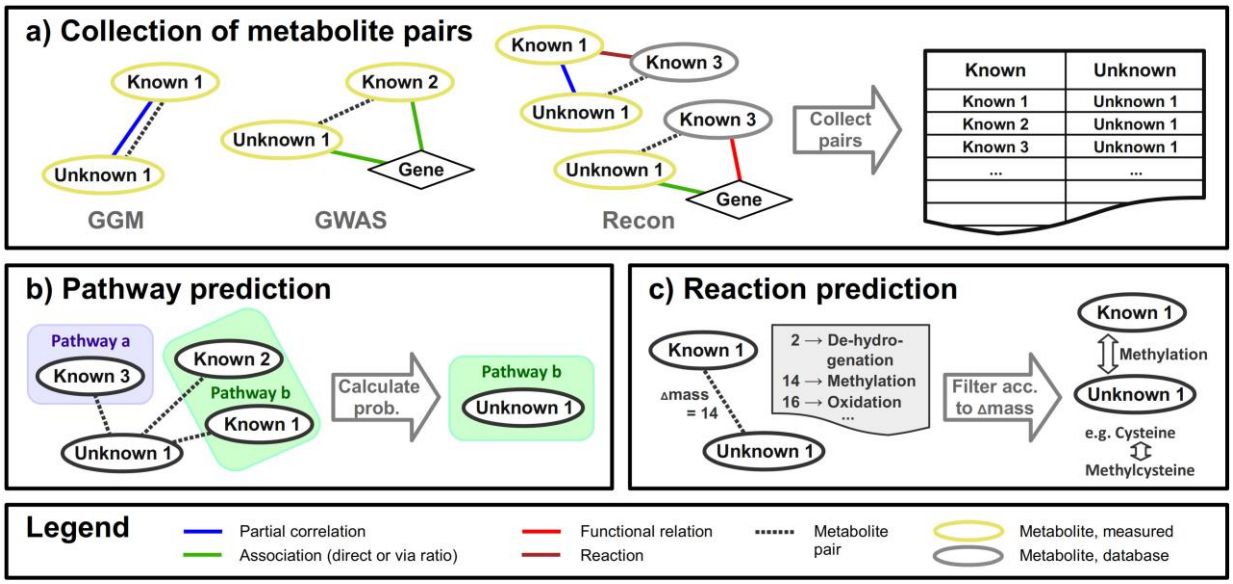


Figure 2: Schema of pathway and reaction prediction
For the prediction of pathways and reactions of unknown metabolites, (a) neighbors of each metabolite were collected based on direct partial correlation (GGM) edges, common genetic links by a GWAS association or functional connections via a Recon metabolite 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 prediction of reactions (c) reactions are assigned between neighboring metabolites based on comparison of their mass difference Δm with a list of mass differences that are characteristic for specific reactions. Note, 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.

3.3 Prediction of reactions

Prediction of reactions, such as methylation, oxidation, hydroxylation, phosphorylation, carboxylation, hydrogenation, etc. (Table 2), between known and unknown metabolites enables the *in silico* application of reactions to known metabolites to select concrete candidates for unknown metabolites.

We tried the simple approach of assigning reactions to pairs of neighboring metabolites in the network which we assumed to be connected by a reaction (Figure 2c) 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. [13] showed that the accuracy of reaction prediction significantly depends on mass resolution, this interval should be adjusted for platforms with better resolution to allow for improved differentiation of reactions. Table 2 shows a summary of pairs of known, known/unknown, and unknown metabolites with assigned reactions. Supplementary Table S3 contains a complete list of assigned reactions per unknown metabolite. We predicted reactions also between pairs of unknown metabolites as it can serve as one

element of metabolite characterization. Pairs among known metabolites were used for verification. 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 2).

Reaction	Δmass	known-known (true*)	known-unknown	unknown-unknown
total pairs		**5600**	**899**	**223**
(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, or Hydroxylation, or Epoxidation	16	75 (48)	14	4
(De)ethylation, or Alkyl-chain-elongation	28	64 (37)	5	7
Quinone, or 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

*) Brackets indicate the number of formally true reactions.

Table 2: Summary of assigned reactions based on Δmass
Frequently occurring reactions are shown with their typical change in mass and their occurrence in the network model among pairs of metabolites ($\Delta m \pm 0.3$). Pairs of metabolites were built based on their neighborhood in network model via GGM, GWAS or Recon edges. In the column of reactions between pairs of neighboring known metabolites, the number of verified reactions among these pairs is indicated.

3.4 Selection of candidate molecules

For a proof of concept, we sought to test our predictions experimentally for the most frequent predicted reaction type, the dehydrogenation reaction. To select the best candidate, we additionally applied a second prediction criterion considering the retention times of the reactants. To this end, we used the distribution of differences in retention index between pairs of known metabolites that are connected by a dehydrogenation reaction compared to known metabolites with the same Δm but connected via another reaction. Out of 15 pairs of connected unknown and known metabolites with $\Delta m = 2 \pm 0.3$, we classified 12 pairs to be part of a dehydrogenation based on this additional criterion (Table S4). 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. Table 3 provides details about five pairs of known and unknown metabolites, classified as lipid (fatty acid: dicarboxylate, medium chain, long chain) with a predicted dehydrogenation reaction.

For experimental validation we focused on the 5 unknown metabolites that were predicted to be fatty acid derivatives (Table 3). In case of these unknown metabolites, 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, 39 molecules remained as concrete candidates for the five unknown metabolites. For 4 out of 5, at least one of the candidate molecules was commercially available.

Metabolite	Super pathway*	Sub pathway*	Reaction*	Reactant*	Candidate molecules	Verified molecule
X-13891	Lipid	Fatty acid, dicarboxylate	dehydrogenation	dodecanedioic acid	**2-dodecendioic acid**, 3-dodecendioic acid, 4-dodecendioic acid, 5-dodecendioic acid, 6-dodecendioic acid	2-dodecendioic acid
X-13069	Lipid	Long chain fatty acid	dehydrogenation	5,8-tetradecadienoate	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	9-tetradecenoic acid
X-11538	Lipid	Fatty acid, dicarboxylate	dehydrogenation	octadecanedioate	2-octadecenedioic acid, 3-octadecenedioic acid, 4-octadecenedioic acid, 5-octadecenedioic acid, 6-octadecenedioic acid, 7-octadecenedioic acid, 8-octadecenedioic acid, **9-octadecenedioic acid**	
X-11859	Lipid	Medium chain fatty acid	dehydrogenation	pelargonate	**2-nonenoic acid**, **3-nonenoic acid**, 4-nonenoic acid, 5-nonenoic acid, 6-nonenoic acid, 7-nonenoic acid, **8-nonenoic acid**	
X-11905	Lipid	Fatty acid, dicarboxylate	dehydrogenation	hexadecanedioate	2-hexadecenedioic acid, 3-hexadecenedioic acid, 4-hexadecenedioic acid, 5-hexadecenedioic acid, 6-hexadecenedioic acid, 7-hexadecenedioic acid, 8-hexadecenedioic acid	

*: automatically predicted features

Table 3: Preselected candidate molecules
We selected candidate molecules for 5 unknown metabolites that we predicted to be fatty acid derivatives and reactants in a dehydrogenation reaction. The structure of the candidate molecules per unknown metabolite basically varies in the position of the predicted double bond. Candidate molecules printed in bold were commercially available and forwarded to the experimental validation.

3.5 Experimental validation of predicted candidate molecules

We bought 6 pure substances of 4 predicted candidates that were available at chemical distributors. Applying LC-MS negative measurements with these pure substances, we were able to verify the predicted identity of two unknown metabolites.

2-dodecendioic acid and X-13891 (m/z: 227.1) share a retention time peak at 2.77 min. in their extracted ion chromatograms (EIC) and show the same fragments with equivalent relative intensities in their MS^2 fragmentation spectra, consequently the candidate molecule is verified (Figure 3).

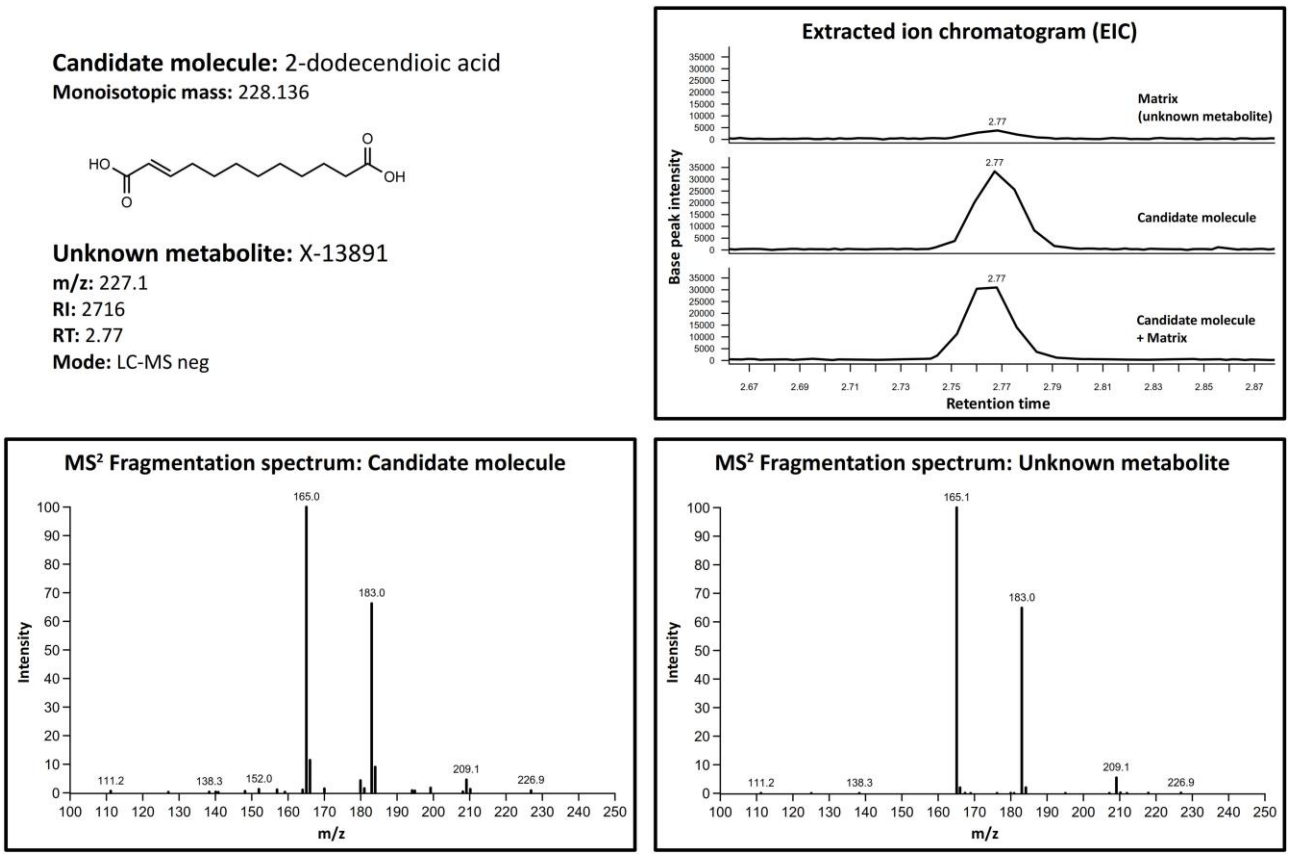


Figure 3: Spectra of the candidate 2-dodecendioic acid and X-13891
The extracted ion chromatograms show the same retention time for each measurement: candidate molecule, candidate molecule + reference matrix (containing the unknown metabolite) and reference matrix. The MS^2 fragmentation spectra of the candidate molecule and of the unknown metabolite show the same fragments with equal relative intensities, consequently the candidate molecule is verified.

Furthermore, we could verify the candidate molecule 9-tetradecenoic acid for X-13069 (Figure S4). 9-octadecenedioic acid and X-11538 show a slight shift in their retention time peaks so that this candidate molecule could be falsified, although both molecules show very similar fragments in their MS^2

fragmentation spectra (Figure S5). 2-nonenoic acid, 3-noneoic acid, 8-noneoic acid and X-11859 have different retention time peaks in the EIC so that these candidates could also be falsified (Figure S6). A detailed overview of the evaluation results of all selected candidate molecules is provided in Table S5.

4 Discussion

In recent years, *in silico* selection of candidates for unknown metabolites became a valuable approach for metabolite identification in non-targeted MS-based metabolomics [5]. Here, we present a new method that uses biochemical and genetic links of unknown and known metabolites rather than their chemical properties derived from spectra to select candidate molecules. Our approach consists of three basic steps: 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 build networks that (i) connect measured metabolites if they show a significant partial correlation forming a Gaussian Graphical Model (GGM) and (ii) connect metabolites to genes if a significant association exists between the metabolite and genetic variation in the gene as derived from a genome-wide association study (GWAS). This approach was previously shown to reconstruct known metabolic pathways [10,23]. In a second step, we integrate substrate, product and genetic information on known biochemical reactions. In our approach we derived this information from Recon2 [12]. While in principle other metabolic databases such as KEGG [21] or HumanCyc [22] can also be used as resources for known biochemical reactions our method does not include parsers for these databases currently. In a third step, we predict pathways and reactions of unknown metabolites based on their neighbors in the network. To assess the quality of these predictions, we evaluated our approach based on the predictions that our method produced for the set of metabolite pairs with known chemical structure.

To test the performance of our approach, we applied it to unknown metabolites from a non-targeted metabolomics platform that was used to determine the blood metabolomes of 2279 subjects. Out of 319 unknown metabolites we were able to characterize 200 metabolites through their neighborhood in the network. For 180 and 109 metabolites we were able to predict pathways and reactions, respectively. Finally,

as a proof of principle, we confirmed our predicted candidates for two unknown metabolites experimentally (X-13891 as 2-dodecendioic acid; X-13069 as 9-tetradecenoic acid) out of four, for which we tested candidates.

Our method complements existing methods in various ways. First, by focusing on the quantitative information across all measured samples as well as genetic and functional associations of metabolites, we make use of orthogonal information in the metabolomics data that is typically omitted in existing approaches, which rely on information from fragmentation spectra (e.g. CFM [7], MetFrag [8]), the calculation of retention time [8] or use networks that depict the similarity of unknown and known metabolites or signals in terms of detected fragments [31], measured mass-to-charge ratios [9,13] or elemental composition [32]. By combining the information that is imprinted in the correlation structure of the data as accessible through our network 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 our case of MS data with low mass resolution. While low mass resolution is insufficient to identify specific mass differences that are typical for certain reactions [13], our network allows pre-filtering these pairs by focusing on neighboring metabolites, which can be assumed to be linked functionally. Second, if metabolomics measurements were performed by companies in a fee-for-service manner (e.g. for large sample sizes), the in depth spectral information about unknown metabolites that is needed for most existing methods is usually not reported. In these cases, our method provides an alternative route for metabolite identification as it does not require spectral details beyond the reported quantities (step 1 and 2) and the mass-to-charge ratio in step 3. Finally, while in methods that do not rely on networks, one run of the method is needed for each unknown metabolite, network-based methods such as the one proposed here provide characterizations for the unknown metabolites in a metabolomics data set in a single run. If new data sets from the same metabolomics platform become available, these data can be easily integrated to an existing network to improve metabolite identification also for the previous data set.

Limitations and future perspectives: One of the major limitations of our approach is its dependence 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. The construction of GGMs requires an at least balanced number of samples and measured parameters. In addition, metabolites with missing values cannot be incorporated into the GGM in our method. In part, we solve this problem through imputation of the missing values. Nonetheless, our method cannot identify candidate molecular structures for unknown metabolites that show a very high number of missing values across the samples, since imputation is not applicable in these cases. Also, if a metabolite is not connected to any known metabolite or gene in our final networks, we cannot provide any further characterization for the unknown metabolite. In general, our approach presupposes a similar behavior of unknown and known metabolites, which ignores potential biases in the distribution of unknown metabolites (e.g. if complete classes of metabolites are unknown). Moreover, in cases where the number of unknown metabolites by far exceeds the number of known metabolites the probability that an unknown metabolite is connected to a known metabolite is low. Thus, a transfer of pathway and reaction information to the unknown molecule is not possible. Although our approach relies on probabilistic graphical models (GGMs), it is not fully probabilistic in the aspects concerning data integration and predictions. A future statistically rigorous extension of our method could incorporate information about the previously identified edges by setting informative priors on GGM structures (using, for example, efficient Bayesian methods for GGMs [33]), or by setting penalties in a regularization-based approach [34]; the latter method could also potentially be used for exclusion of hubs or unknown confounders. Future inference of the sub- and super-pathways can potentially be based on less stringent independence assumptions. A further limitation of our approach is the current usage of a very simplified method for reaction prediction by solely relying on mass differences. In addition, the currently tested list of possible enzymatic reactions is not complete. This step can be improved by applying chemometric methods for in silico reaction prediction that take functional groups of the known metabolite into account [35].

5 Conclusion

Metabolite identification for non-targeted MS-based platforms is one of the major bottlenecks of metabolomics approaches today [5]. Here we described a method that uses biochemical and genetic information as imprinted in the correlation structure of the measured metabolite levels and genotype-metabotype links from genome-wide association studies to select candidate molecules for unknown metabolites. Integration of these metabolite-metabolite and metabolite-gene pairs with functional data from known enzymatic reactions into a network embeds unknown metabolites into their context of metabolic pathways. Predicting pathways and reactions of unknown metabolites based on their neighborhood to known metabolites in the network thereby allows identification of possible candidates. Combining our approach with methods that use orthogonal chemical information on unknown metabolites such as fragmentation or isotope patterns will largely improve metabolite identification in future studies.

Acknowledgements

This project was in part supported with funding from the Innovative Medicines Initiative within the "Surrogate markers for Micro- and Macro-vascular hard endpoints for Innovative diabetes Tools" (SUMMIT) project. WRM was supported by the Helmholtz Cross Program Initiative "Personalized Medicine (iMed)". The authors would like to thank Bianca Eichner and Anna Artati of the Genome Analysis Center, Helmholtz Zentrum München for support in the use of LC-MS during the verification of candidate molecules.

References

[1] T. Fuhrer, High-throughput discovery metabolomics, Curr. Opin. Biotechnol. 31 (2015) 73–78. doi:10.1016/j.copbio.2014.08.006.

[2] R.D. Beger, W. Dunn, M.A. Schmidt, S.S. Gross, J.A. Kirwan, M. Cascante, L. Brennan, D.S. Wishart, M. Oresic, T. Hankemeier, D.I. Broadhurst, A.N. Lane, K. Suhre, G. Kastenmuller, S.J. Sumner, I. Thiele, O. Fiehn, R. Kaddurah-Daouk, Metabolomics enables precision medicine: "A White Paper, Community Perspective"., Metabolomics. 12 (2016) 149. doi:10.1007/s11306-016-1094-6.

[3] P. Yin, G. Xu, Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications, J. Chromatogr. A. 1374 (2014) 1–13.

doi:http://dx.doi.org/10.1016/j.chroma.2014.11.050.

[4] L.W. Sumner, Z. Lei, B.J. Nikolau, K. Saito, U. Roessner, R. Trengove, Proposed quantitative and alphanumeric metabolite identification metrics, Metabolomics. 10 (2014) 1047–1049. doi:10.1007/s11306-014-0739-6.

[5] D.J. Creek, W.B. Dunn, O. Fiehn, J.L. Griffin, R.D. Hall, Z. Lei, R. Mistrik, S. Neumann, E.L. Schymanski, L.W. Sumner, R.D. Trengove, J.-L. Wolfender, Metabolite identification: Are you sure? And how do your peers gauge your confidence?, Metabolomics. 10 (2014) 350–353. doi:10.1007/s11306-014-0656-8.

[6] E. Werner, J.-F. Heilier, C. Ducruix, E. Ezan, C. Junot, J.-C. Tabet, Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends., J. Chromatogr. B. 871 (2008) 143–63. doi:10.1016/j.jchromb.2008.07.004.

[7] F. Allen, R. Greiner, D. Wishart, Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, Metabolomics. 11 (2015) 98–110. doi:10.1007/s11306-014-0676-4.

[8] C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender, S. Neumann, MetFrag relaunched: incorporating strategies beyond in silico fragmentation, J. Cheminform. 8 (2016) 1–16. doi:10.1186/s13321-016-0115-9.

[9] D. Grapov, K. Wanichthanarak, O. Fiehn, MetaMapR: pathway independent metabolomic network analysis incorporating unknowns, Bioinformatics. 31 (2015) 2757–2760. doi:10.1093/bioinformatics/btv194.

[10] J. Krumsiek, K. Suhre, A.M. Evans, M.W. Mitchell, R.P. Mohney, M. V Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski, C. Gieger, F.J. Theis, G. Kastenmüller, Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information, PLoS Genet. 8 (2012) 1–14. doi:10.1371/journal.pgen.1003005.

[11] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, F.J. Theis, Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data, BMC Syst. Biol. 5 (2011) 1–16. doi:10.1186/1752-0509-5-21.

[12] I. Thiele, N. Swainston, R.M.T. Fleming, A. Hoppe, S. Sahoo, M.K. Aurich, H. Haraldsdottir, M.L. Mo, O. Rolfsson, M.D. Stobbe, S.G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A.K. Chavali, P. Dobson, W.B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J.J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J.A. Papin, N.D. Price, E. Selkov, M.I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J.H.G.M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V Westerhoff, D.B. Kell, P. Mendes, B.Ø. Palsson, A community-driven global reconstruction of human metabolism, Nat. Biotechnol. 31 (2013) 419–425. doi:10.1038/nbt.2488.

[13] R. Breitling, S. Ritchie, D. Goodenowe, M.L. Stewart, M.P. Barrett, Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data, Metabolomics. 2 (2006) 155–164. doi:10.1007/s11306-006-0029-z.

[14] K. Borodulin, E. Vartiainen, M. Peltonen, P. Jousilahti, A. Juolevi, T. Laatikainen, S. Männistö, V. Salomaa, J. Sundvall, P. Puska, Forty-year trends in cardiovascular risk factors in Finland, Eur. J. Public Health. 25 (2015) 539–546. doi:10.1093/eurpub/cku174.

[15] E.R. Pearson, L.A. Donnelly, C. Kimber, A. Whitley, A.S.F. Doney, M.I. McCarthy, A.T. Hattersley, A.D. Morris, C.N.A. Palmer, Variation in TCF7L2 Influences Therapeutic Response to Sulfonylureas: a GoDARTs study, Diabetes. 56 (2007) 2178–2182. doi:10.2337/db07-0440.

[16] D. Baldassarre, K. Nyyssönen, R. Rauramaa, U. de Faire, A. Hamsten, A.J. Smit, E. Mannarino, S.E. Humphries, P. Giral, E. Grossi, F. Veglia, R. Paoletti, E. Tremoli, Cross-sectional analysis of baseline data to identify the major determinants of carotid intima-media thickness in a European population: the IMPROVE study, Eur. Heart J. 31 (2010) 614–622. doi:10.1093/eurheartj/ehp496.

[17] A.C. Carlsson, P.E. Wandell, G. Journath, U. de Faire, M.-L. Hellenius, Factors associated with uncontrolled hypertension and cardiovascular risk in hypertensive 60-year-old men and women a population-based study, Hypertens Res. 32 (2009) 780–785. http://dx.doi.org/10.1038/hr.2009.94.

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

[19] A.M. Evans, C.D. DeHaven, T. Barrett, M. Mitchell, E. 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, Anal. Chem. 81 (2009) 6656–6667. doi:10.1021/ac901536h.

[20] K.J. Boudonck, M.W. Mitchell, J. Wulff, J.A. Ryals, Characterization of the biochemical variability of bovine milk using metabolomics, Metabolomics. 5 (2009) 375–386. doi:10.1007/s11306-009-0160-8.

[21] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation, Nucleic Acids Res. 44 (2016) D457–D462. doi:10.1093/nar/gkv1070.

[22] P. Romero, J. Wagg, M.L. Green, D. Kaiser, M. Krummenacker, P.D. Karp, Computational prediction of human metabolic pathways from the complete human genome, Genome Biol. 6 (2004) R2. doi:10.1186/gb-2004-6-1-r2.

[23] S.-Y. Shin, E.B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A.M. Valdes, C.L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, T.M.T.H.E.R. (MuTHER) Consortium, M. Waldenberger, J.B. Richards, R.P. Mohney, M. V Milburn, S.L. John, J. Trimmer, F.J. Theis, J.P. Overington, K. Suhre, M.J. Brosnan, C. Gieger, G. Kastenmuller, T.D. Spector, N. Soranzo, An atlas of genetic influences on human blood metabolites, Nat Genet. 46 (2014) 543–550. http://dx.doi.org/10.1038/ng.2982.

[24] S.S. Chen, R.A. Gopinath, Gaussianization, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Adv. Neural Inf. Process. Syst. 13, MIT Press, 2001: pp. 423–429. http://papers.nips.cc/paper/1856-gaussianization.pdf.

[25] R Core Team, R: A Language and Environment for Statistical Computing, (2015). https://www.r-project.org/.

[26] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, J. Stat. Softw. 45 (2011) 1–67. http://www.jstatsoft.org/v45/i03/.

[27] J. Schaefer, R. Opgen-Rhein, K. Strimmer., GeneNet: Modeling and Inferring Gene Networks, (2015). http://cran.r-project.org/package=GeneNet.

[28] W.S. van Helden J, Wernisch L, Gilbert D, Graph-based analysis of metabolic networks, Ernst Scher. Res Found Work. 38 (2002).

[29] V. Lacroix, L. Cottret, P. Thébault, M.F. Sagot, An Introduction to Metabolic Networks and Their Structural Analysis, IEEE/ACM Trans. Comput. Biol. Bioinforma. 5 (2008) 594–617. doi:10.1109/TCBB.2008.79.

[30] C. Frainay, F. Jourdan, Computational methods to identify metabolic sub-networks based on metabolomic profiles, Brief. Bioinform. 18 (2017) 43. doi:10.1093/bib/bbv115.

[31] J.J.J. van der Hooft, J. Wandy, M.P. Barrett, K.E. V Burgess, S. Rogers, Topic modeling for untargeted substructure exploration in metabolomics, Proc. Natl. Acad. Sci. . (2016). doi:10.1073/pnas.1608041113.

[32] S. Kim, R.P. Rodgers, A.G. Marshall, Truly “exact” mass: Elemental composition can be determined uniquely from molecular mass measurement at ~0.1 mDa accuracy for molecules up to ~500 Da, Int. J. Mass Spectrom. 251 (2006) 260–265. doi:http://dx.doi.org/10.1016/j.ijms.2006.02.001.

[33] P. Orchard, F. Agakov, A. Storkey, Bayesian Inference in Sparse Gaussian Graphical Models, ArXiv E-Prints. (2013).

[34] F. V Agakov, P. Orchard, A.J. Storkey, Discriminative Mixtures of Sparse Latent Fields for Risk Management, in: N.D. Lawrence, M.A. Girolami (Eds.), J. Mach. Learn. Res. - Work. Conf. Proc., 2012: pp. 10–18. http://jmlr.csail.mit.edu/proceedings/papers/v22/agakov12/agakov12.pdf.

[35] R. Höllering, J. Gasteiger, L. Steinhauer, K.-P. Schulz, A. Herwig, Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis, J. Chem. Inf. Comput. Sci. 40 (2000) 482–494. doi:10.1021/ci990433p.

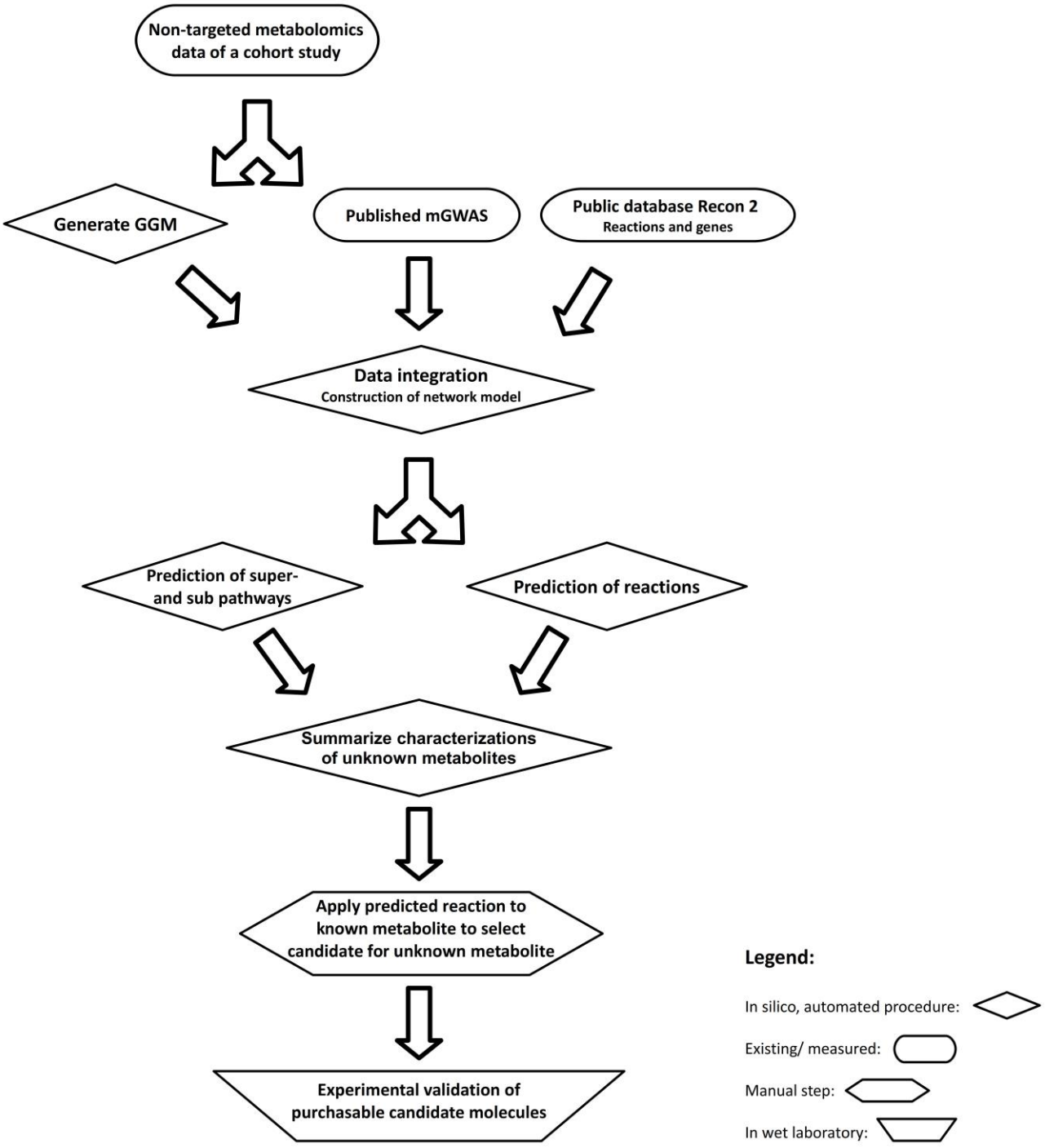


Figure S1: Workflow of the complete procedure
The schematic workflow shows all steps of our procedure to characterize unknown metabolites. The shape of each element indicates the respective part consisting of an automated *in silico* procedure or existing data or manual or wet laboratory work.

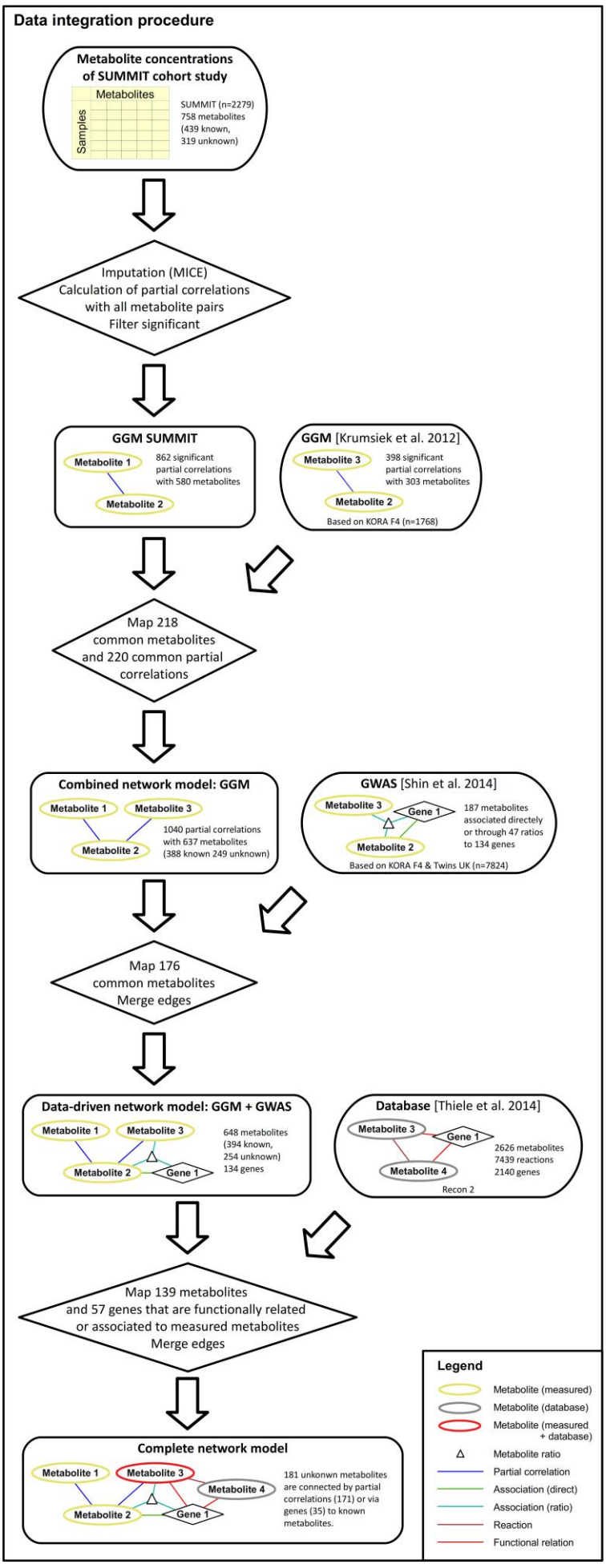


Figure S2: Data integration procedure
The network model is composed by three different types of data, which are integrated in separate steps of data integration. The workflow describes each step together with its input and output, which represent intermediate states of the network.

Confidence[1]	Total	Correct	Similar	False	Ff[2]	None
very high	10797	8354 (77.4%)	953 (8.8%)	1317 (12.1%)	183 (1.7%)	0 (0.0%)
high	1278	682 (53.4%)	0 (0.0%)	533 (41.7%)	63 (4.9%)	0 (0.0%)
medium	9914	5932 (59.8%)	1302 (13.1%)	1781 (18.0%)	899 (9.1%)	0 (0.0%)
low	5308	1957 (36.9%)	93 (1.8%)	1941 (36.6%)	1218 (22.9%)	99 (1.9%)
very low	2985	827 (27.7%)	305 (10.2%)	691 (23.1%)	1162 (38.9%)	0 (0.0%)

1: Confidence of super pathway prediction, 2) false because of wrong super pathway prediction

Table S1: Cross validation of the pathway prediction module

A 10-fold cross validation was applied for 100 times to get the total count of predicted super pathways per confidence level. The 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, or no prediction.

Name	Predicted super pathway	Confidence	Predicted sub pathway	Fraction[1]
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
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
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
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
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

1: Fraction of the most frequent sub pathway among neighboring metabolites with the predicted super pathway

Table S2: Predicted super and sub pathways of unknown metabolites

The table contains a list of all predicted super pathways for the set of unknown metabolites including their confidence classes ordered by decreasing confidence. The predicted sub pathways are shown with its fraction in surrounding metabolites with the predicted super pathway.

Metabolite 1	Metabolite 2	Super Pathway 2	Sub Pathway 2	Δmass	Predicted reaction by Δmass
11-HpODE	X-11421			0.98	Oxidative deamination
2-methylbutyroylcarnitine	X-11255	amino acid	valine, leucine and isoleucine metabolism	1*	Oxidative deamination
pyroglutamine	X-11315	amino acid	glutamate metabolism	1*	Oxidative deamination
3-methylhistidine	X-13835	amino acid	histidine metabolism	1*	Oxidative deamination
X-14632	X-14658			1	Oxidative deamination
Guanine	X-11422			1.05	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	-1.92	De-hydrogenation/ Reduction
decanoylcarnitine	X-13435	lipid	carnitine metabolism	-1.94	De-hydrogenation/ Reduction
estrone	X-02249			-1.96	De-hydrogenation/ Reduction
pseudouridine	X-11429	nucleotide	pyrimidine metabolism,	2*	De-hydrogenation/ Reduction

			uracil containing		
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	X-11469	lipid	fatty acid, dicarboxylate	-2*	De-hydrogenation/ Reduction
X-11299	X-11483			-2*	De-hydrogenation/ Reduction
octadecanedioate	X-11538	lipid	fatty acid, dicarboxylate	-2*	De-hydrogenation/ Reduction
hexadecanedioate	X-11905	lipid	fatty acid, dicarboxylate	-2*	De-hydrogenation/ Reduction
thymol sulfate	X-12847	xenobiotics	food component/plant	-2*	De-hydrogenation/ Reduction
X-11538	X-16480			-2	De-hydrogenation/ Reduction
L-urobilin	X-17162	cofactors and vitamins	hemoglobin and porphyrin metabolism	-2*	De-hydrogenation/ Reduction
X-12844	X-17357			2	De-hydrogenation/ Reduction
X-12846	X-17703			-2	De-hydrogenation/ Reduction
X-15492	X-17706			-2	De-hydrogenation/ Reduction
omega hydroxy tetradecanoate (n-C14:0)	X-11438			-2	De-hydrogenation/ Reduction
pelargonate (9:0)	X-17269	lipid	medium chain fatty acid	-2.02	De-hydrogenation/ Reduction
dehydroisoandrosterone sulfate (DHEA-S)	X-18601	lipid	sterol, steroid	2.04	De-hydrogenation/ Reduction
4-hydroxyphenylpyruvate	X-12543	amino acid	phenylalanine & tyrosine metabolism	2.07	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	Methylation/ De-methylation, or Alkyl-chain-elongation
X-11317	X-11497			14	Methylation/ De-methylation, or Alkyl-chain-elongation
13-cis-retinoate	X-11530			14	Methylation/ De-methylation, or Alkyl-chain-elongation
all-trans-retinoate	X-11530			14	Methylation/ De-methylation, or Alkyl-chain-elongation
X-14374	X-14473			14	Methylation/ De-methylation, or Alkyl-chain-elongation
X-12212	X-15636			14	Methylation/ De-methylation, or Alkyl-chain-elongation
X-11441	X-16946			-14	Methylation/ De-methylation, or Alkyl-chain-elongation
X-12734	X-17685			14	Methylation/ De-methylation, or Alkyl-chain-elongation
palmitate (16:0)	X-11438	lipid	long chain fatty acid	-14.03	Methylation/ De-methylation, or Alkyl-chain-elongation
estrone	X-02269			-14.06	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	-15.93	Oxidation, or Hydroxylation, or Epoxidation
3-carboxy-4-methyl-5-propyl-2-furanpropanoate (CMPF)	X-02269	lipid	fatty acid, dicarboxylate	16*	Oxidation, or Hydroxylation, or Epoxidation
p-cresol sulfate	X-06126	amino acid	phenylalanine &	16*	Oxidation, or Hydroxylation,

			tyrosine metabolism		or Epoxidation
tetradecanedioate	X-11438	lipid	fatty acid, dicarboxylate	-16*	Oxidation, or Hydroxylation, or Epoxidation
X-11444	X-11470			-16	Oxidation, or Hydroxylation, or Epoxidation
4-ethylphenylsulfate	X-12230	xenobiotics	benzoate metabolism	16*	Oxidation, or Hydroxylation, or Epoxidation
X-12329	X-12730			16	Oxidation, or Hydroxylation, or Epoxidation
X-12217	X-12734			16	Oxidation, or Hydroxylation, or Epoxidation
2-aminooctanoic acid	X-13477	amino acid	amino fatty acid	16*	Oxidation, or Hydroxylation, or Epoxidation
gamma-glutamylglutamate	X-14272	peptide	gamma-glutamyl amino acid	-16*	Oxidation, or Hydroxylation, or Epoxidation
catechol sulfate	X-16940	xenobiotics	benzoate metabolism	16*	Oxidation, or Hydroxylation, or Epoxidation
hypoxanthine	X-11422	nucleotide	purine metabolism, (hypo)xanthine/inosine containing	16.06	Oxidation, or Hydroxylation, or Epoxidation
estradiol	X-02269			-16.08	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	-27.83	De-ethylation, or Alkyl-chain-elongation
docosapentaenoate (n3 DPA; 22:5n3)	X-12627	lipid	essential fatty acid	28*	De-ethylation, or Alkyl-chain-elongation
X-12855	X-12860			28	De-ethylation, or Alkyl-chain-elongation
X-16674	X-17438			28	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*	Quinone, or CH3 to COOH, or Nitro reduction
omega hydroxy hexadecanoate (n-C16:0)	X-11438			-30.03	Quinone, or CH3 to COOH, or Nitro reduction

16alpha-Hydroxyestrone	X-02269			-30.06	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
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)	-31.93	Bis-oxidation
pregnenolone sulfate	X-12456	lipid	sterol/steroid	32.01	Bis-oxidation
estrone	X-11469			-32.06	Bis-oxidation
2-Hydroxyestradiol-17beta	X-02269			-32.07	Bis-oxidation
linolenate [alpha or gamma; (18:3n3 or 6)]	X-16480	lipid	essential fatty acid	32.08	Bis-oxidation
lathosterol	X-12063	lipid	sterol/steroid	41.85	Acetylation
lathosterol	X-12456	lipid	sterol/steroid	41.85	Acetylation
5alpha-cholest-8-en-3beta-ol	X-12063			41.85	Acetylation
5alpha-cholest-8-en-3beta-ol	X-12456			41.85	Acetylation
androsterone	X-11441			41.88	Acetylation
androsterone	X-11442			41.88	Acetylation
carnosine	X-11561			41.99	Acetylation
X-12253	X-12258			42	Acetylation
2-aminooctanoic acid	X-12511	amino acid	amino fatty acid	42*	Acetylation
X-12802	X-12860			-42	Acetylation
Isocitrate	X-15245			42	Acetylation
estradiol	X-11530			42.02	Acetylation
laurate	X-11438			42.03	Acetylation
1-docosahexaenoylglycerophosphocholine	X-12644	lipid	lysolipid	-42.1*	Acetylation
aflatoxin B1 exo-8,9-epoxide	X-18601			42.14	Acetylation
cholesta-5,7-dien-3beta-ol	X-12063			43.86	Decarboxylation
cholesta-5,7-dien-3beta-ol	X-12456			43.86	Decarboxylation
5alpha-cholesta-7,24-dien-3beta-ol	X-12063			43.86	Decarboxylation
5alpha-cholesta-7,24-dien-3beta-ol	X-12456			43.86	Decarboxylation
testosterone	X-11441			43.89	Decarboxylation
testosterone	X-11442			43.89	Decarboxylation
glucose	X-12007	carbohydrate	glycolysis, gluconeogenesis, pyruvate metabolism	43.94	Decarboxylation
hexadecanedioate	X-11438	lipid	fatty acid, dicarboxylate	-44*	Decarboxylation
andro steroid monosulfate 2	X-12063	lipid	sterol/steroid	44*	Decarboxylation
oxalatosuccinate(3-)	X-15245			44.01	Decarboxylation
estrone	X-11530			44.04	Decarboxylation
acetylcarnitine	X-12465	lipid	carnitine metabolism	44.08	Decarboxylation
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,17beta-diol disulfate 2	X-18601	lipid	sterol, steroid	79.98	Sulfation, or Phosphatation
X-06126	X-11837			80	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*	Sulfation, or Phosphatation
alpha-glutamyltyrosine	X-11805	peptide	dipeptide	107*	Taurine Conjugation
X-12830	X-17703			107.1	Taurine Conjugation
X-11261	X-11478			-119	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	-120.86	Cys Conjugation
testosterone sulfate	X-11440			-120.86	Cys Conjugation
deoxycholate	X-11491	lipid	bile acid metabolism	176.02	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
phenylalanylphenylalanine	X-17189	peptide	dipeptide	306.8*	GSH Conjugation

*) Δmass based on measured mass of known metabolite

Table S3: Predicted reactions based on Δmass

We selected reactions connecting known and unknown metabolites as well as among pairs of unknown metabolites simply based on Δmass. Pairs of metabolites were collected based on GGM edges, a common GWAS gene, or a gene that is functionally related to a metabolite and associated to an unknown metabolite. The sign of the numbers of the Δmass column indicate the direction of the reaction, starting with the known metabolite (or metabolite 1).

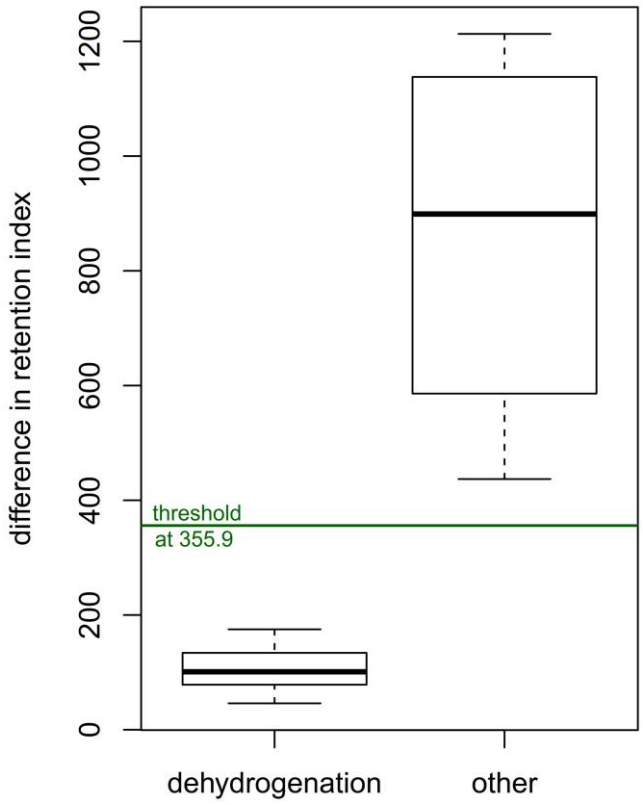


Figure S3: Difference in retention index of metabolites connected by a dehydrogenation reaction

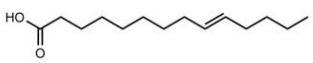
The distributions of difference in retention time of pairs of known metabolites connected by a GGM edge are shown separately for reaction partners of dehydrogenation reactions or other cases. The threshold was calculated by the mean of the distributions $\pm$ their standard deviations.

Metabolite 1	Metabolite 2	Δmass	Δri	Predicted reaction
pelargonate (9:0)	X-11859	-1.92	294	De-hydrogenation/ Reduction
decanoylcarnitine	X-13435	-1.94	60	De-hydrogenation/ Reduction
pseudouridine	X-11429	2	47	De-hydrogenation/ Reduction
octadecanedioate	X-11538	-2	113	De-hydrogenation/ Reduction
hexadecanedioate	X-11905	-2	249	De-hydrogenation/ Reduction
thymol sulfate	X-12847	-2	155	De-hydrogenation/ Reduction
L-urobilin	X-17162	-2	50.2	De-hydrogenation/ Reduction
pelargonate (9:0)	X-17269	-2.02	316.8	De-hydrogenation/ Reduction
dehydroisoandrosterone sulfate (DHEA-S)	X-18601	2.04	229.4	De-hydrogenation/ Reduction
4-hydroxyphenylpyruvate	X-12543	2.07	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	NA	measured on different platforms
estrone	X-02249	-1.96	NA	non-measured metabolite
omega hydroxy tetradecanoate (n-C14:0)	X-11438	-2	NA	non-measured metabolite
X-11444	X-12844	-1.9	185	De-hydrogenation/ Reduction
X-12230	X-17185	-1.9	290.9	De-hydrogenation/ Reduction
X-11538	X-16480	-2	234.5	De-hydrogenation/ Reduction
X-12844	X-17357	2	73	De-hydrogenation/ Reduction
X-12846	X-17703	-2	70	De-hydrogenation/ Reduction
X-15492	X-17706	-2	6.7	De-hydrogenation/ Reduction
X-17359	X-17706	-2.1	83.8	De-hydrogenation/ Reduction
X-11378	X-16935	1.8	859.5	no
X-12217	X-16940	1.8	648.9	no
X-11444	X-17706	-1.9	708.3	no
X-11299	X-11483	-2	450	no

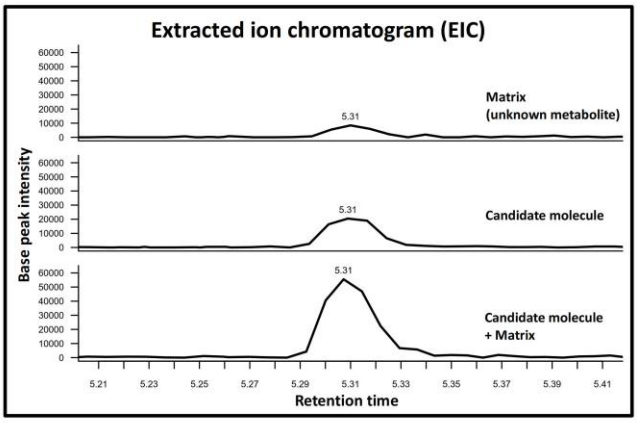
Table S4: Predicted dehydrogenation/ reduction reactions

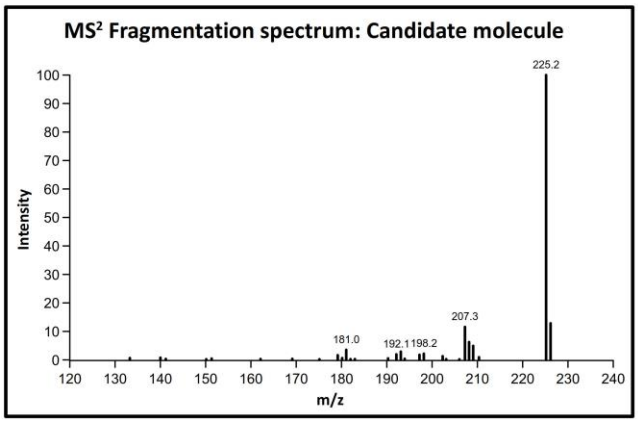
The table contains all pairs of neighboring known and unknown metabolites with mass difference 2 ± 0.3 and a prediction whether or not they are connected by a dehydrogenation reaction. Additionally, pairs of unknown metabolites are also considered. The sign of the numbers of the column Δmass indicate the direction of the reaction, starting with the known metabolite or metabolite 1. The numbers in column Δri contain absolute values.

Candidate molecule: 9-tetradecenoic acid
Monoisotopic mass: 226.193



Unknown metabolite: X-13069
m/z: 225.4
RI: 5380
RT: 5.31
Mode: LC-MS neg





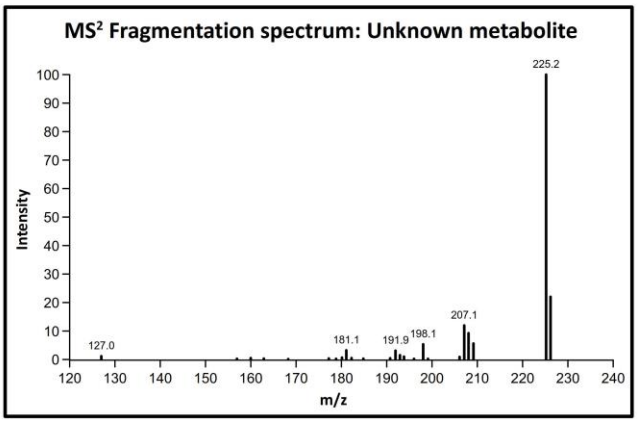


Figure S4: Spectra of the candidate molecule 9-tetradecenoic acid and X-13069
The extracted ion chromatograms show the same retention time peaks for the candidate molecule, the unknown metabolite, and the mixture of both substances. Their MS^2 fragmentation spectra are composed of the same fragments with equal relative intensities. Therefore, this candidate is verified.

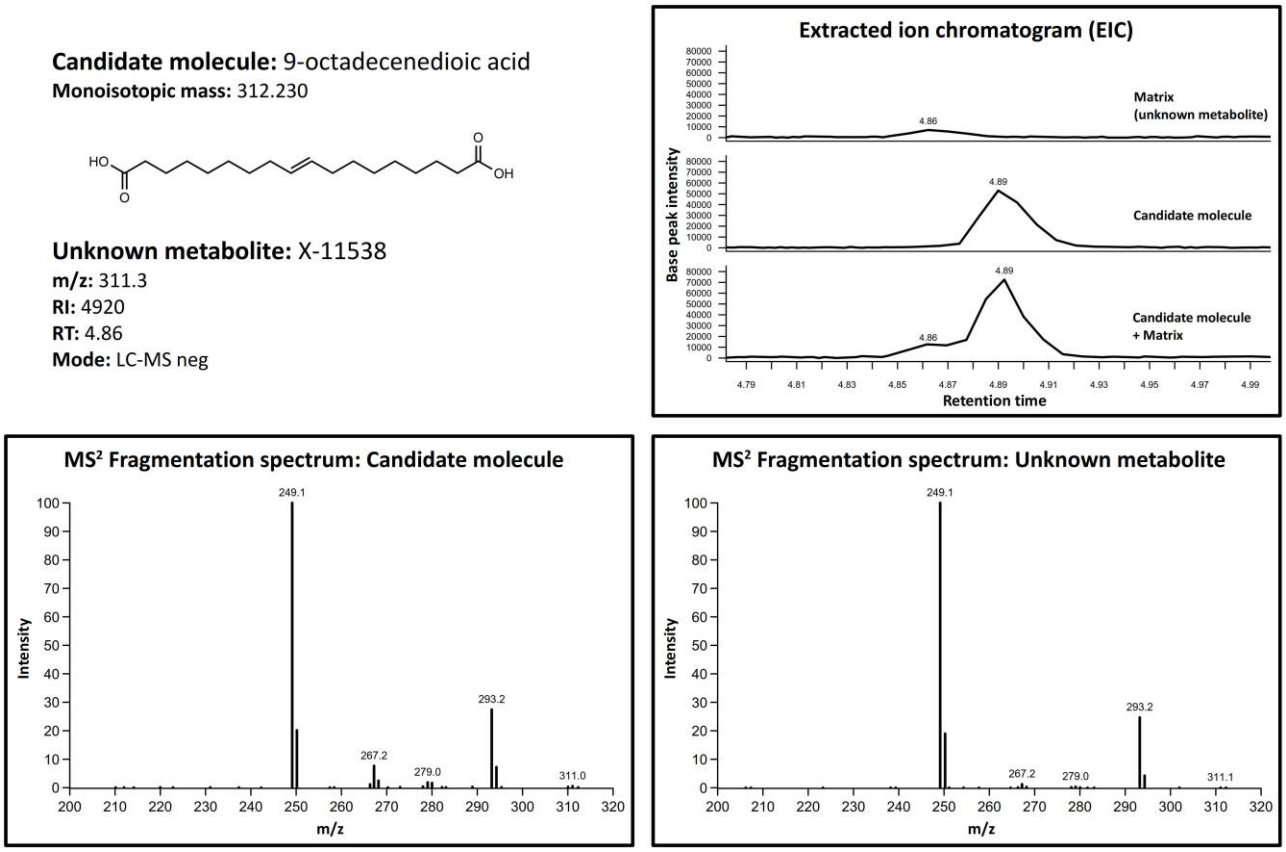


Figure S5: Spectra of the candidate molecule 9-octadecenedioic acid and X-11538
The retention time peaks of 9-octadecenedioic acid and X-11538 are slightly shifted (4.89 and 4.86 min.), which can be seen in the extracted ion chromatogram of the mixture probe of the pure substance and the reference plasma that contains the unknown molecule. The fragmentation spectra are similar, but not identical. The two main peaks at 249.1 and 293.2 show similar relative intensities in the MS^2 fragmentation spectra of all measurements, but two smaller peaks at 267.2 and 279.1 are much larger for the plasma sample compared to the pure substance. In consequence this candidate could be falsified.

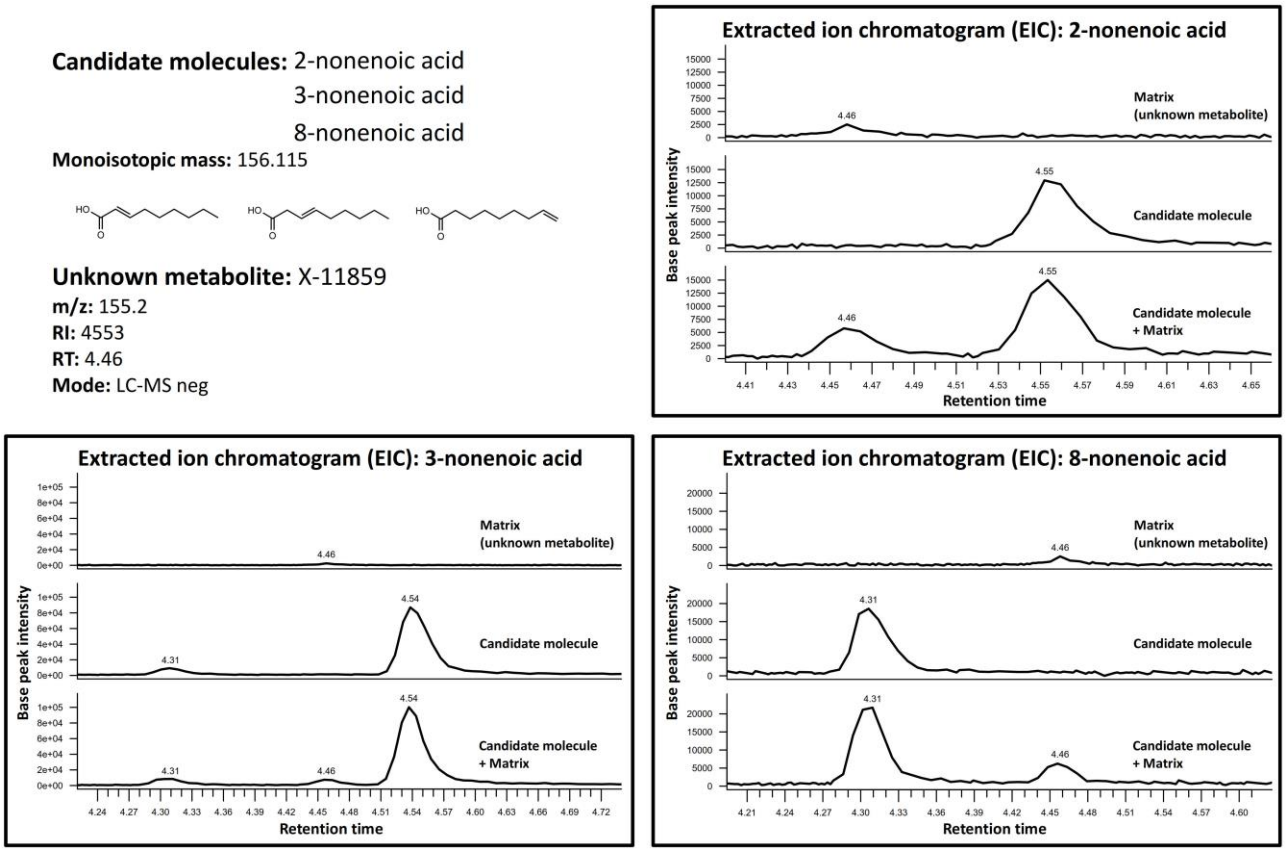


Figure S6: Extracted ion chromatograms of the candidate molecules of X-11859
2-nonenoic acid, 3-noneoic acid, 8-noneoic acid and X-11859 (m/z: 155.2) have different retention time peaks in the extracted ion chromatograms. Partially, there are fragments with the same m/z in the MS^2 fragmentation spectra, but the relative intensities are different so that these candidates could be falsified.

Compound 1	Compound 2	m/z	RT 1	RT 2	MS^2 1	MS^2 2	match
X-13891	2-dodecendioic 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 4.31	110.8, 82.0, 155.2, 123.1, 136.9	111.1, 155.0, 137.1 155.0, 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 S5: Detailed evaluation results of predicted candidate molecules
The table juxtaposes the retention time and the main MS^2 fragments, ordered by descending intensity, of the purchased candidate molecules and the respective unknown metabolites of LC-MS measurements. Both characteristics match for 2 selected candidate molecules and differs in 4 cases.

File S1: R scripts and example data
The zip file contains implementations for all modules of the approach in R along with example data.

File S2: Graphml file of a representation of the network model

The network model embeds 254 measured unknown metabolites into their biochemical and functional context. The model was constructed based on 758 measured metabolites (known: 439, unknown: 319), 2626 metabolites of the public database Recon 2 and 1782 genes. For clarity, elements of Recon 2 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 1.