Supporting Information for

Solutions for low versus high resolution mass spectrometric data fusion. A data-driven annotation strategy in non-targeted metabolomics

Sara Forcisi, [*,†,⊥,#] *Franco Moritz,* [†,⊥,#] *Marianna Lucio,* [†] *Rainer Lehmann,* [§,∥,⊥] *Norbert Stefan,* [∥,⊥,×] *and Philippe Schmitt-Kopplin* [*,†,‡,⊥]

*These authors contributed equally.

[†] Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research

Center for Environment Health, Neuherberg, D-85764, Germany

[‡] Chair of Analytical Food Chemistry, Technische Universität München, Freising-Weihenstephan, D- 85354, Germany

[§] Division of Clinical Chemistry and Pathobiochemistry (Central Laboratory), University Hospital

Tübingen, D-72076, Germany

‖ Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Centre Munich at the University of Tübingen (IDM), D-72076 Tübingen, Germany

⊥ German Center for Diabetes Research (DZD), Germany

× Department of Internal Medicine, Division of Endocrinology, Diabetology, Nephrology, Vascular Disease and Clinical Chemistry, University Hospital of the Eberhard Karls University, D-72076 Tübingen, Germany

*To whom correspondence should be addressed: sara-forcisi@helmholtz-muenchen.de and schmitt-kopplin@helmholtz-muenchen.de

Tables describing the NAFL population and the TOF-MS parameters are given. Details pertaining to the experimental section are described: instrumental and methodological parameters, patient descriptions, as well an introduction into error plots and isobar plots. Supporting results (Figures) as referenced in the main text are given. Supporting results pertain to an error distribution and an isobar plot, visualizations of statistical modelling and compound classes most affected by them as well as *m/z* over RT plots of the positioning of statistically relevant compound classes.

Table S1: NAFL population characteristics. The 40 individuals are separated according to their insulin sensitivity ($ISI_{Matsuda}$).

Characteristics	**Insulin sensitive subjects**	**Insulin resistant subjects**	**p value**
Insulin sensitivity [ISI-Matsuda Index]	14.7 ± 1.1	7.3 ± 0.52	< 0.0001
Gender [males/females]	15/5	7/13	0.02
Age [years]	52 ± 2	44 ± 2	0.016

Table S2: MS detection parameters applied in UHPLC-QTOF-MS experiment.

Detection Parameters	Positive ESI
Capillary Voltage (kV)	3.10
Sample Cone (V)	30.00
Extraction cone (V)	4.00
Source Temperature (ºC)	120
Desolvatation temperature (ºC)	300
Desolvatation Flow (L/h)	800
Cone(L/h)	50
Detector (V)	1800

EXPERIMENTAL SECTION

Chemicals. Methanol (LC/MS grade) and water (LC/MS grade) were purchased from Fluka Analytical (Sigma-Aldrich), H_3PO_4 85 % was purchased from Merck, formic acid (LC/MS grade) was purchased from Fluka Analytical (Sigma-Aldrich). Leucine Enkephalin solution (400 ng/µl) was purchased from Waters (Milford, USA). [d_3]acetyl-L-carnitine.HCl and [16,16,16-d_3]hexadecanoyl-L-carnitine.HCl were purchased from Dr. Herman J. ten Brink, VU medical center, Netherlands. [d_{10}]adipic acid, nialamide, [d_6]sulfadimethoxine, reserpine, [d_4]cholic acid and decanoic Acid-C13 were purchased from Sigma-Aldrich. A description of the solutions and samples used in UHPLC-QTOF-MS experiment is given in the supplementary material.

Patients. The population of 40 patients with NAFLD consisted of 75% males in the insulin sensitive (IS) group and 65% females in the insulin resistant (IR) group. Following the explanation of the nature and possible consequences of the studies from all participants informed written consent was obtained. The local medical ethics committee had approved the protocol. Insulin sensitivity was calculated from a frequently sampled oral glucose tolerance test (oGTT) [1] and mirrors insulin sensitivity in the muscle and in the liver showing lower values in the prediabetic state. The results presented in this work, consider the separation of the population according to the insulin sensitivity recorded at baseline before subjects underwent a lifestyle intervention (Tübingen Lifestyle Intervention Program (TULIP)). In

table S1 (see supplementary material) the insulin sensitivity characteristics of the population are depicted [2].

Plasma sample collection and treatment. EDTA blood was collected from the 40 patients with NAFLD participating TULIP after an overnight fast. The precise phenotypical characterization of the subjects was previously reported [2]. Blood plasma was prepared at once via centrifugation at 2,000 g at 4 ºC for 7 min. The resulting plasma was aliquoted for storage at -80 ºC. Prior to the MS analyses the frozen EDTA plasma samples were thawed on ice and vortex-mixed for 30 seconds prior to further treatment.

For FT-MS analyses a C18 Solid phase extraction (SPE) technology using Omix C18 100µl tips (Varian) was adopted for metabolite extraction. An aliquot of 50µl of plasma was diluted (1:1) in 2% H_3PO_4 and vortex mixed for 30 seconds before loading it onto the SPE tip. The vendor guide line was followed for the sample preparation procedure. For the conditioning and equilibration, MeOH and 2% formic acid were used. For the washing step, formic acid 2% was used and MeOH for the elution. The eluate was diluted (dil. factor: 1/50) in MeoH /H2O:8/2 prior FT-ICR/MS analysis. Prior to UHPLC-MS analyses, a protein precipitation extraction (PPE) was performed via addition of ice-cold acetonitrile (320 µl) to a plasma volume (80 µl). The samples were vortex-mixed for 30 s at room temperature and centrifuged at 15,294 g for 10 min at 4 ºC. The samples were freeze-dried and the pellet was the reconstituted in 50 µl of 20% acetonitrile; QC plasma samples were reconstituted in 50 µl of standard solution (see supplementary material).

Solutions and samples used in UHPLC-QTOF-MS experiment

MS

Calibration Solution Prepare a solution of sodium formate mixing 10 %Vol. of formic acid in water (1 part), 0.1 M sodium hydroxide in water (1 part) and acetonitrile (8 part)

Lock Mass Solution Add 0.2 % Vol. of Leucine Enkephaline solution and 0.1 % Vol. of formic acid to a solution of MeOH/H_2O:1/1 (LC-MS grade) to a final concentration of 400 µg/l. This reference compound is injected at regular intervals of 15 seconds from an orthogonal spray (the lock mass spray).

UHPLC

Mobile Phases

All the solvents are LC/MS grade:

A: H_2O/AcN : 95/5 + 0.1 % formic acid

B: AcN

weak wash: H_2O/AcN : 9/1

strong wash: AcN/ MeOH/2-propanol : 6/3/1

Samples

Blank

H_2O/AcN: 80/20.

Plasma Blank

Water collected in plasma Sarstedt tubes and treated as plasma; then reconstituted in H_2O/AcN : 80/20 at volume of 50 µl.

Plasma QC

Pool of plasma from different donors treated as plasma and reconstituted in 50 µl of the standard solution.

Standard solution

A mixture of 5 standards (each standard was prepared as stock solution of 1mg/ml) was prepared in a solution of H2O/AcN: 8/2. The mixed solution was used to reconstitute QC-plasma samples after protein precipitation. The volume of reconstitution was 50 µl. In this way each standard was diluted, reaching the final concentration of 1 mg/L (Table S3) per standard. In table S4 the standard compounds used during the experiments are described in terms of sum formula, monoisotopic mass (*m/z*), the possible detectable positive ion adducts and the relative retention time (RT) of the standards.

Table S3: Final concentration of internal standards.

Nr.	Name	STOCK SOLUTION 1 mg/ml standard solved in	RECONSTITUTION (Final Concentration (in a final volume of 50 µl of H_2O/AcN:8/2) mg/L
1	[d_3]Acetyl-L-carnitine HCl	Methanol	1
2	Nialamide	Acetonitrile	1
3	[d_6]Sulfadimethoxine	Acetonitrile	1
4	Reserpine	Acetonitrile	1
5	[16,16,16d_3]hexadecanoyl-L carnitine HCl	Methanol	1

Table S4: Isotope labeled standard compounds used in NAFL study detected in positive ionization mode.

Nr.	Name	Formula	Monoisotopic Ion mass (*m/z*)	Possible ion adduct ions (*m/z*)			RT (min)
				$[M+H]^+$	$[M+Na]^+$	$[M+K]^+$	
1	[d_3]Acetyl-L-Carnitine	$C_9D_3H_{14}NO_4$	207.141865	207.14186	229.12381	245.09774	1.27±0.2
2	Nialamide	$C_{16}H_{18}N_4O_2$	298.142976	299.15052	321.13219	331.10613	5.60±0.2
3	[d_6]Sulfadimethoxine	$C_{12}D_6H_8N_4O_4S$	316.111236	317.11851	339.10045	355.07439	8.30±0.2
4	Reserpine	$C_{33}H_{40}N_2O_9$	608.273381	609.28065	631.26260	647.23653	10.61±0.2
5	[d_3]Hexadecanoyl-L-Carnitine	$C_{23}D_3H_{42}NO_4$	403.361514	403.36151	425.34345	441.31739	16.56±0.2

Error Distributions. Error over *m/z* distributions (EMZDs) characterize the performance of a mass spectrometer at any measured *m/z*. Currently, they are poorly used in mass spectrometry literature; the major reason probably being the predominance of targeted mass spectrometry. EMZDs are of high importance for the evaluation of sum formula annotations, as the center of the EMZD can shift towards positive or negative errors locally. Such shifts can be caused by high ion abundances. In contrast to TOF instruments, FT-ICR/MS has a 'memory' for such peak events in that an abnormally large ion cloud of given *m/z* alters the center of the EMZD for higher *m/z* regions. Knowing the center of an error distribution at any *m/z* is of major importance: if a mass spectrometer has a precise mass shift towards e.g. +1 ppm, then any formula calculation that minimizes the error towards that position will be false by default. We therefore propose the use of EMZDs for the evaluation of broad band annotation quality (supplementary information). It is to be noted, that EMZDs that are produced on the basis of combinatorial formula assignment under minimization of error will always be centered on 0 ppm. However, on this basis it is not possible to evaluate whether the finding is true or not unless sufficient ion abundances allow for the detection of isotopologues and/or for MS/MS. This is not the case for the vast majority of features. MDiN-based formula annotation bypasses this problem as it walks 'along' the intrinsic EMZD of a mass spectrum as long as the starting masses were assigned correctly.

Isobar plots. A tool for the evaluation of the goodness of MDiN based formula assignment is the isobar plot. In MDiN based formula assignments all annotations are dependent on another. Depending on the starting masses, overall mass accuracy and search direction, an *m/z* peak may attain different annotations as the annotation procedure is repeated. If an annotation is repeated e.g. 100 times, *m/z* peaks will attain no annotation, one (constant) annotation or multiple annotations. These results can be displayed in a bar chart, the isobar plot (supplementary information). A symmetric isobar distribution within the isobar plot indicates entirely random formula assignment within the bounds of a given error window (Figure S2 or Figure 1B). If the isobar plot has a median isobaric formula assignment of one formula throughout all annotation cycles with multiple assignments describing an exponentially decreasing tail, the underlying network structure is non-random and there is little dissent between a node's formula assignment and its REMDs.

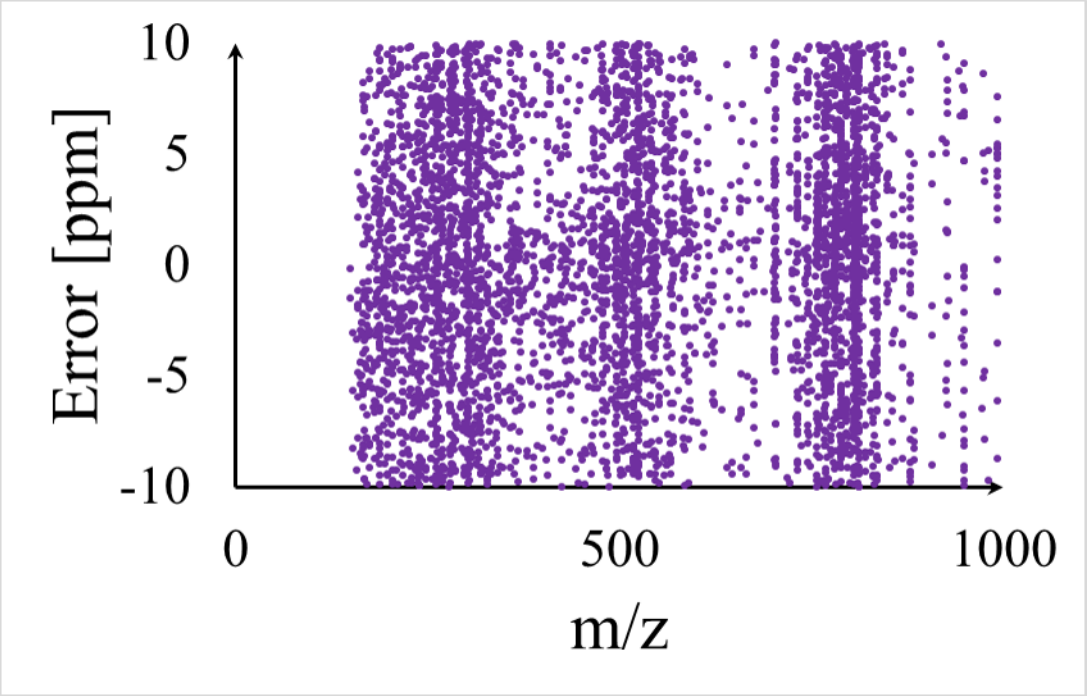

Figure S1: Error over *m/z* distribution of matching experimental UHPLC-MS data onto ICR-FT/MS data at an error of ± 10 ppm.

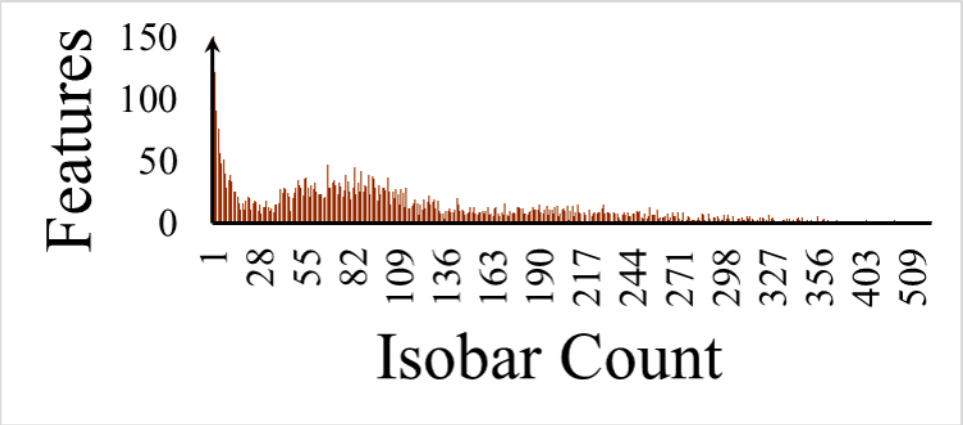

Figure S2: Distribution of isobaric annotation counts per LC-MS feature. Formulas were calculated by a combinatorial in-house written software (Formcalc). The error window was set to ± 2 ppm and the maximum formula was set to $C_{100}O_{70}N_{20}P_3S_3Na$. Formulas were filtered for the Senior rules.

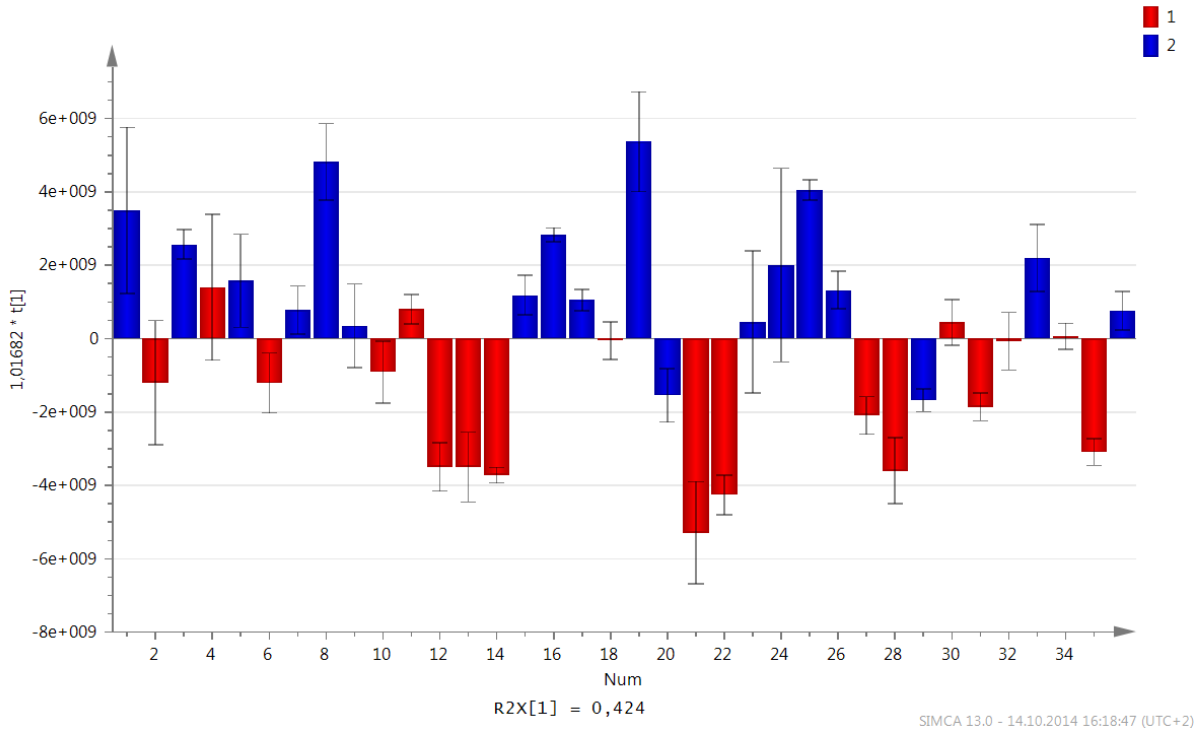

Figure S3: OPLS score scatter plot in which is visualized the two different groups [blue= insulin sensitive (IS) and red= insulin resistant (IR)], with $R^2Y(cum)=0.5$ and $Q^2(cum)=0.4$. CV Anova with $p<0.05$.

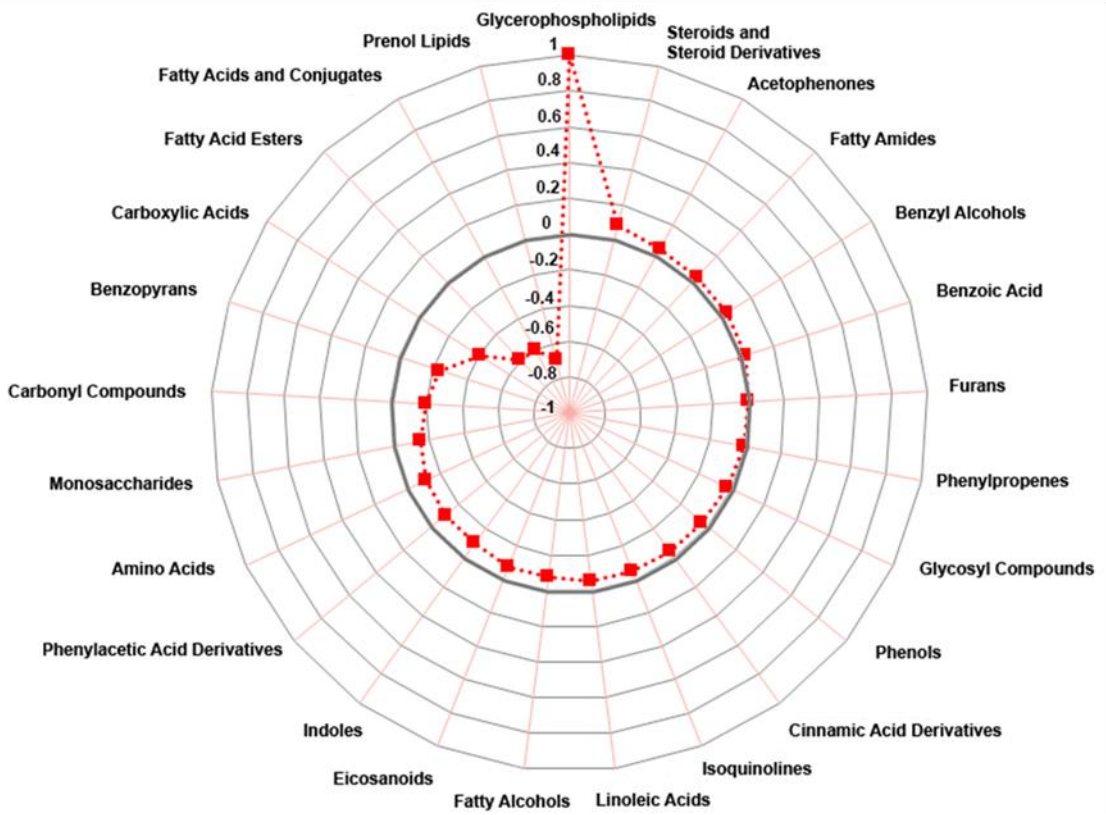

Figure S4: Radar plot of FT-ICR/MS compound classes. Centered and normalized rank sums indicate their general location on the multivariate statistic. Positive values indicate a prevalence of annotations in IS and negative values indicate prevalence in IR.

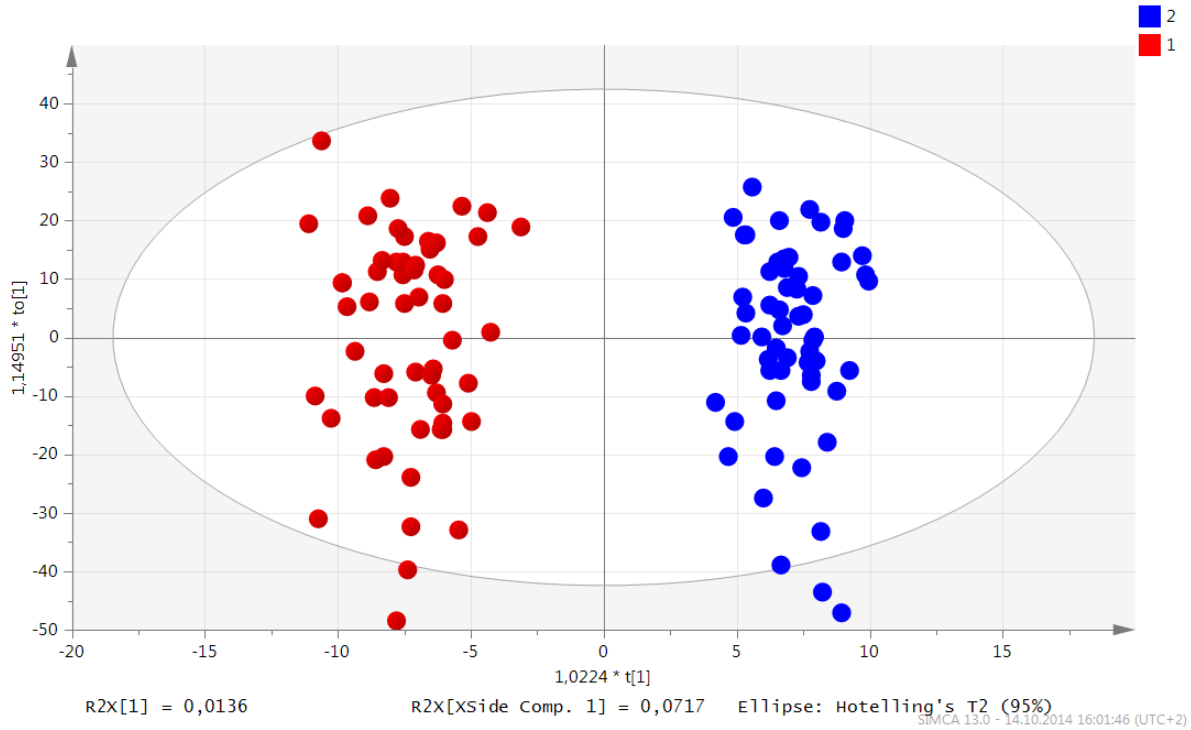

Figure S5: OPLS scatter plot describing the separation among the two classes of interest (IR=1 and IS=2) of the population of fatty liver subjects. The study shows a good model fit ($R^2Y(cum)=0.9$) and sufficient productiveness in cross-validation ($Q^2(cum)=0.4$). CV Anova with $p<0.001$.

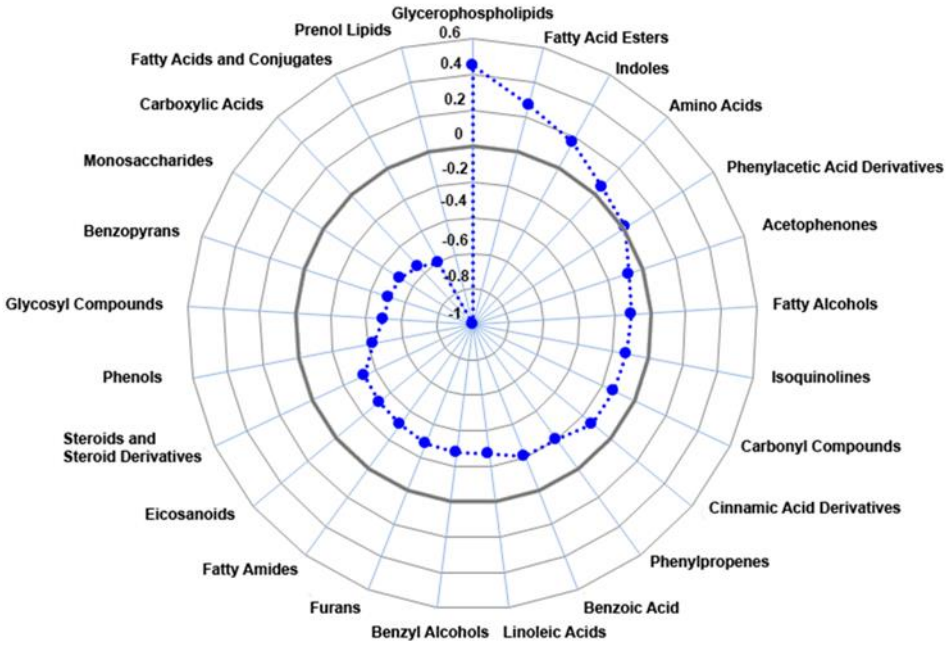

Figure S6: Radar plot of RP-UHPLC-QTOF-MS compound classes. Centered and normalized rank sums indicate their general location on the multivariate statistic. Positive values indicate a prevalence of annotations in IS and negative values indicate prevalence in IR.

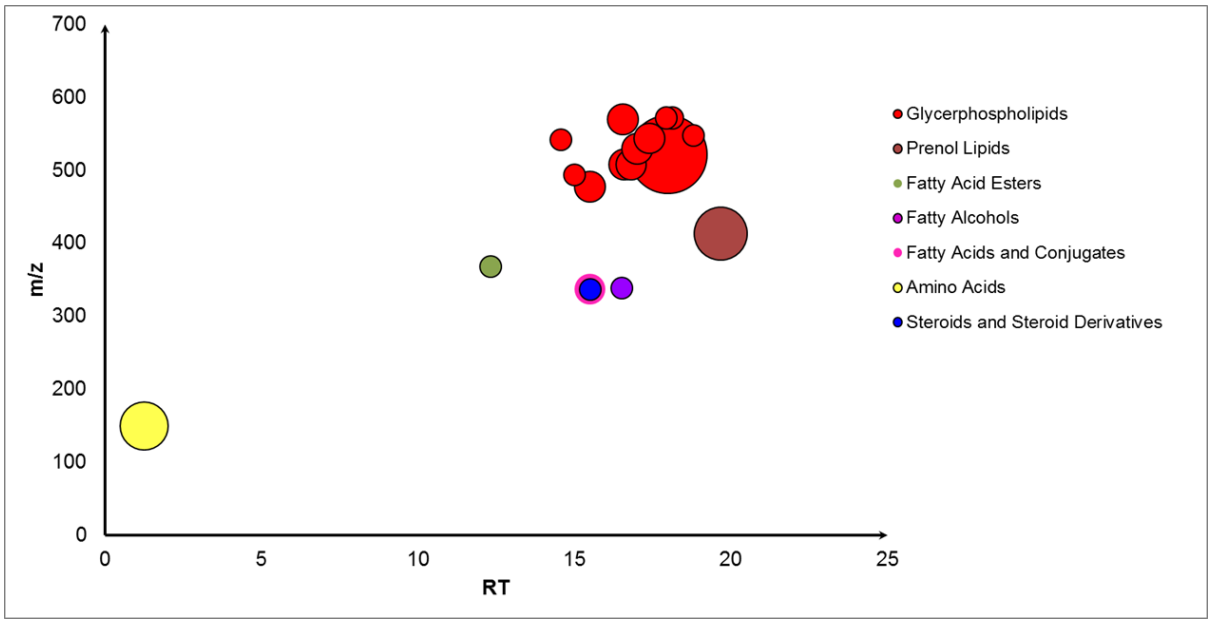

Figure S7: Up-regulated classes of compounds (in IS class) as function of the retention time (RT). Coinciding points indicate isomers that are not differentiable by RP-LC-MS.

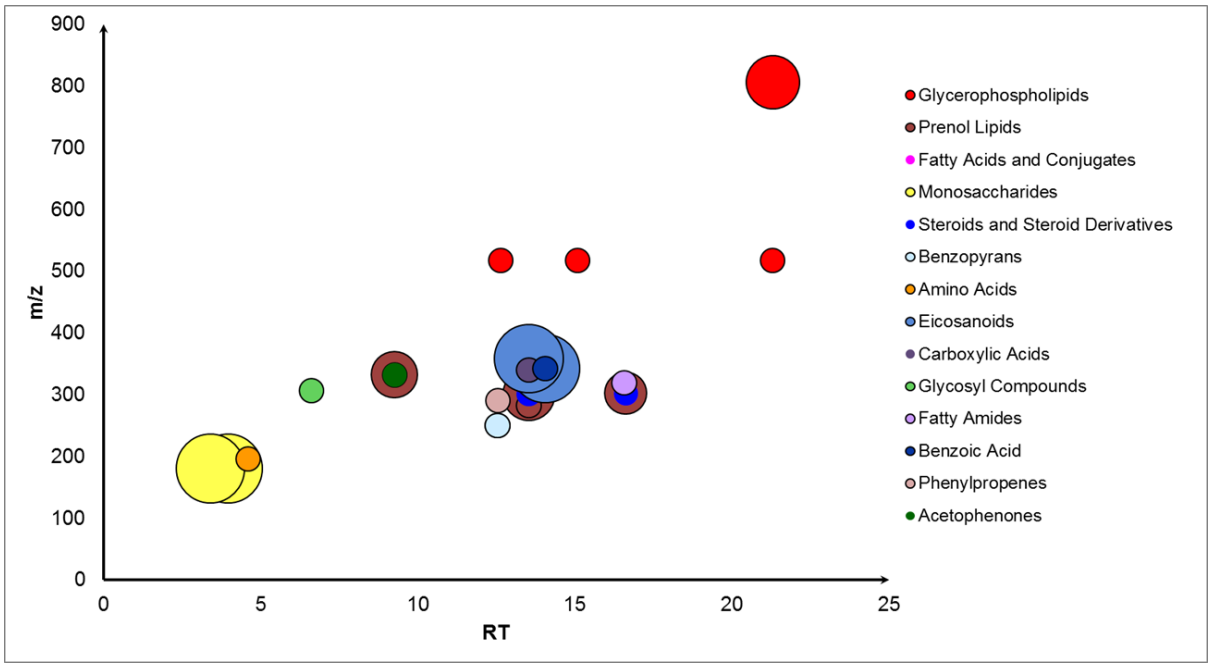

Figure S8: Down-regulated classes of compounds (in IS class) as function of the retention time (RT). Coinciding points indicate isomers that are not differentiable by RP-LC-MS.

[1] M. Matsuda, R.A. DeFronzo, Diabetes Care 22 (1999) 1462.
[2] R. Lehmann, H. Franken, S. Dammeier, L. Rosenbaum, K. Kantartzis, A. Peter, A. Zell, P. Adam, J. Li, G. Xu, A. Konigsrainer, J. Machann, F. Schick, M. Hrabe de Angelis, M. Schwab, H. Staiger, E. Schleicher, A. Gastaldelli, A. Fritsche, H.U. Haring, N. Stefan, Diabetes Care 36 (2013) 2331.