analytical chemistry

pubs.acs.org/ac

Perspective

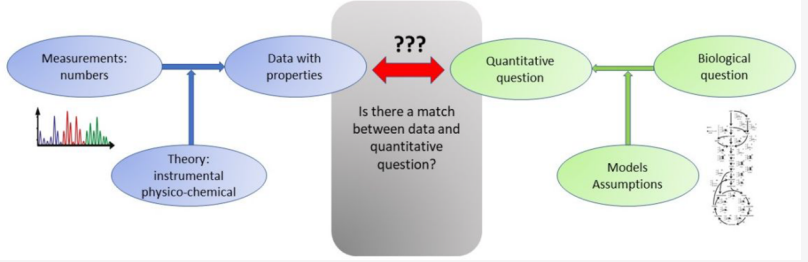
Numerical Representations of Metabolic Systems

Age K. Smilde* and Thomas Hankemeier

Cite This: *Anal. Chem.* 2020, 92, 13614−13621

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Metabolomics is becoming a mature part of analytical chemistry as evidenced by the growing number of publications and attendees of international conferences dedicated to this topic. Yet, a systematic treatment of the fundamental structure and properties of metabolomics data is lagging behind. We want to fill this gap by introducing two fundamental theories concerning metabolomics data: data theory and measurement theory. Our approach is to ask simple questions, the answers of which require applying these theories to metabolomics. We show that we can distinguish at least four different levels of metabolomics data with different properties and warn against confusing data with numbers. This treatment provides a theoretical underpinning for preprocessing and postprocessing methods in metabolomics and also argues for a proper match between type of metabolomics data and the biological question to be answered. The approach can be extended to other omics measurements such as proteomics and is thus of relevance for a large analytical chemistry community.

Metabolomics concerns the measurement of small biochemical compounds (metabolites) in samples obtained from biological systems or, in a broader context, from samples that contain such metabolites (extracts from natural foods, environmental samples, etc.). Such measurements are subsequently used to infer relevant information about the associated (biological) system related to a certain research question.

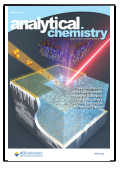
Nowadays, there is a whole variety of metabolomics measurements available which can be categorized either by the type of instruments used (mostly liquid chromatography−mass spectrometry (LC−MS), gas chromatography−mass spectrometry (GC−MS), capillary electrophoresis−mass spectrometry (CE−MS), and NMR) or by the type of measurement performed. The latter pertains to whether the measurement is targeted to a certain number of (known) metabolites or to an untargeted analysis in which also (many) unknown metabolites are being measured. There are also methods which are a combination of both. A typical pipeline for a metabolomics study runs through different steps: formulating a biological question, experimental design, sampling, measuring, preprocessing the data, analyzing the preprocessed data, visualization of results, and answering the biological question.[1]

An often neglected part of the above-mentioned pipeline is the difference between numbers and data. This is a very fundamental issue at the heart of any measurement. We will explain this issue starting from asking a few very simple questions about whether numbers in a metabolomics experiment can be meaningfully compared to each other. To answer these questions, we have to introduce two theories, namely, data theory and measurement theory. After that, we will give (partly) answers to the questions and try to come to a synthesis. As a running example throughout this paper, we will consider measuring lipids in blood using LC−MS.

The goals of this paper are (1) provide a theoretical underpinning of preprocessing methods; (2) give guidelines for a proper use of data analysis methods and propose alternatives; (3) warn against conclusions which are not supported by the (properties of) the data; (4) argue for a proper match between data properties, biological question, and data analysis method; and (5) creating awareness that numbers (from an instrument) is not yet data.

In short, we provide a theoretical framework for thinking about and dealing with metabolomics data. In a broader context, we would like to create awareness that numbers are not data, which is highly relevant in this era of Big Data. We

Received: December 12, 2019
Accepted: September 29, 2020
Published: September 29, 2020



https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621

will not discuss the specifics of the different preprocessing and data analysis methods nor of related topics such as missing data handling and measurement error. There are many papers already discussing this. We invite the metabolomics practitioners to apply our framework on their way of analyzing metabolomics data.

SIMPLE QUESTIONS

We start by visualizing the numbers obtained from an LC−MS experiment on lipids in blood (see Figure 1). The raw data can

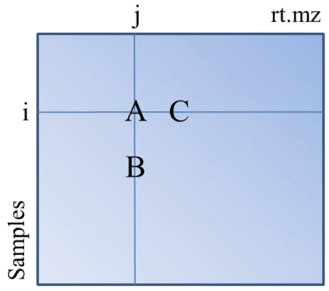


Figure 1. Schematic of raw measurements of lipids in blood. Legend: *i* is a row in the matrix; *j* is a column; *A*, *B*, and *C* are specific numbers in the matrix; rt.mz is the retention time-mass spectrometry index.

be arranged in intensities obtained at a certain m/z value at a certain retention time (rt), and the combined index rt.mz indicates a column in the matrix containing the samples in its rows. Looking at Figure 1, we can ask simple questions to what extent the numbers are comparable, specifically:

(1) If $A > C$; does that have a meaning?
(2) If $A > B$; does that have a meaning?
(3) Does $A - C$ have a meaning?
(4) Does $A - B$ have a meaning?
(5) Does A/C have a meaning?
(6) Does A/B have a meaning?

which should be taken as examples, e.g., when $A < C$, then question one has to change accordingly. Note that moving from question one to three puts a higher demand on the numbers, e.g., if A/C is meaningful then necessarily $A > C$ must have a meaning (but not vice versa!). This notion will be formalized later.

The questions asked above are relevant for a subsequent data analysis. Take the example of PCA (For a short explanation of the methods, see the Supporting Information.), the workhorse of metabolomics data analysis. The score-plots of a PCA are usually interpreted in terms of distances between dots representing the samples, where samples far apart are regarded as very dissimilar and vice versa. However, scores are linear combinations of the original variables (i.e., numbers), and this assumes that for distances in scores plots to be meaningful, at least also the original numbers should be comparable (at least at the level of $A-C$). Similar reasonings hold for loading plots and for the results of OPLS-DA and other often used techniques. Hence, it makes sense to answer the above posed questions.

Actually, there is even a more basic question to ask before considering the simple questions: is it even meaningful to start comparing the values *A*, *B*, and *C*? This question is key in the field known as data theory. The next questions regarding at which level comparisons are possible is the subject of measurement theory. Therefore, both will be explained briefly in the sequel. This paper will be mainly concerned with mass-spectrometry based measurements; the case for NMR is a little different and will be touched upon in the end.

DATA THEORY

A set of notions regarding comparability is called data theory and was pioneered by Coombs[2] and explained for multiway analysis.[3] The first important notion in data theory is conditionality, where we can distinguish column-, row-, and matrix-conditionality.

When considering numbers arranged in a matrix (such as in Figure 1), then different types of comparisons can be made: between numbers across rows in the same column (Figure 1, between *A* and *B*) and between numbers across columns in the same row (Figure 1, between *A* and *C*). When such data can be compared meaningfully, the data are called column-conditional and row-conditional, respectively. When data can be meaningfully compared across rows and columns, then these data are called matrix-conditional.

The prototypical example of row-conditional data are metabolomics measurements of urine, e.g., using NMR. Depending on the different urine histories of the subjects, the urine can be more or less concentrated. This makes the values within one column of a data matrix incomparable since the (unknown) dilution factor of the subjects destroys the comparability. The typical solution of this problem is found in normalizing the different samples thereby attempting to achieve matrix-conditionality. Whether this completely solves the problem is a matter of debate and it also depends on the research question. Actually, different types of metabolites are differently excreted by the kidney: some are only excreted by filtration, some are (partly) readsorbed, and readsorption is achieved by different transporters, for example, one for acidic amino acids, one for dibasic amino acids, one for neutral amino acids.[4] This could justify a normalization per certain metabolite classes rather than normalizing all metabolites in the same manner. Moreover, the type of normalization may also depend on the type of sample, e.g, whether it originates from urine, serum, or tissue. Discussing these issues further is beyond the scope of this paper.

A more serious problem regarding comparability as discussed in data theory is lack-of-invariance: the numbers in a single column do not have the same meaning. This problem is more fundamental than conditionality. Whereas in conditionality, numbers cannot be compared since there are (unknown) arbitrary differences, in lack-of-invariance the meaning of the variables changes within a column. The prototypical example is unsynchronized time series data (see Figure 2, panel a). The time series of three subjects are collected for multiple metabolites; in this figure, only one metabolite is shown. The series are not synchronized, therefore the measurements at, e.g., physical time point t_4 cannot be compared across subjects because they pertain to different states of the biological process measured with the metabolite. Hence, the meaning of the measurement at time point t_4 changes and is not invariant.

A naive arrangement of the numbers is shown in Figure 2, panel b). This is called naive since the lack-of-invariance is not taken into account. A more accurate arrangement of the numbers is shown in Figure 3 because now it is clear that each subject has its own unique time points (see the subscript *i* on the variables indicating time points). Obviously, the numbers as shown in Figure 3 cannot be used as such. Remedies of this

https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621

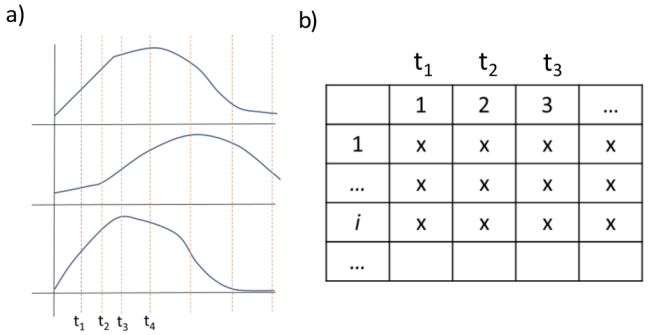


Figure 2. Lack-of-invariance illustrated: (a) unsynchronized times series of several subjects and (b) the naive arrangement of the numbers.

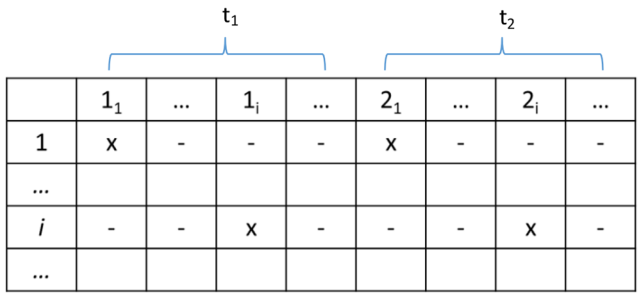


Figure 3. Proper arrangement of the numbers whereby each individual *i* receives its own time points.

problem are found in alignment procedures (e.g., using warping approaches[5]). After such an alignment of all metabolites, assuming that this has solved the lack-of-invariance problem, the numbers can be arranged in a three-way array and analyzed with proper three-way methods such as PARAFAC.[6]

It is important to realize that such lack-of-invariance problems can also be solved by using data analysis methods that do not require the numbers to be synchronized. One such an alternative for the case of Figure 2 is to concatenate the data sets per subject (time versus metabolites) in such a way that all subject-matrices are stacked on top of each other with the metabolites as the common mode. Then methods like simultaneous component analysis (SCA)[7] can be used. This is one of the ways to solve the synchronization problems in batch statistical process monitoring where the batches are also not synchronized.[8] Hence, the properties of the data have repercussions on the methods that can be applied and the type of biological questions that can be solved.

MEASUREMENT THEORY

After having established that a comparison between numbers is meaningful, the next question is at what level this can be done. This was pioneered by Stevens[9] and later taken up and further developed by several authors.[10−13] A nice introduction is given by Hand[14] and a summary is given in Table 1 which is explained briefly. In the Supporting Information, we give a more formal treatment with an illustrative example.

The basic notion in measurement theory is that we want to represent (properties of) a system by numbers, i.e., we want to give a numerical representation of a system. The lowest measurement level is nominal data which are merely (exclusive) categories. Examples are different types of cars,

Table 1. Formal Treatment of Types of Data Scales[a]

scale-type	example	permissible transformations	permissible statistics
nominal	categories	one-to-one	number of cases
ordinal	survey data	monotonic	median, IQR
interval	degree Celsius	positive linear transformation	mean, standard deviation
	calender time	$x' = \alpha x + \beta$ $(\alpha > 0)$	
ratio	length mass	similarity transformation $x' = \alpha x$ $(\alpha > 0)$	coefficient of variation
absolute	counts	$x' = x$	all previous

[a]For explanation, see the text.

different countries, etc. The data are only used as class labels, and these can be changed as long as each class receives a unique other label. Hence, the permissible transformations, the transformations between numerical representations that keep the relationships in the corresponding system intact, are one-to-one transformations. The type of statistics to be used for this type of data are number of cases, frequencies, χ^2-tests, etc.

The next level of measurement scale are ordinal data. The prototypical example is survey data in which respondents can score on certain issues using the answers strongly disagree, disagree, neutral, agree, strongly agree. Obviously, there is an order in these answers; and these answers can be labeled from 1 to 5. The difference between 2 and 1 on the one hand and between 3 and 2 on the other hand, although exactly equal, does not have a meaning. The system can also be represented using a different set of numbers, e.g, 2, 4, 7, 8, 9, but the transformation between the two numerical representations needs to be monotonic. The type of statistics to be employed are the ones for the lower-scaled measurement (i.e., nominal data) and in addition median, interquartile range (IQR), etc.

Interval-scale data is the next level. An example is degree Celsius where the numbers zero and hundred are arbitarily chosen. Stated otherwise, this scale does not have a natural zero and unit. This means that another scale (x') can be used with $x' = \alpha x + \beta$ $(\alpha > 0)$, and this scale has the same meaning for the system; an example is Fahrenheit where $\alpha = 9/5$ and $\beta = 32$. Nevertheless, the ratio of differences between values of this scale has meaning in terms of the system, e.g., in using calendar times $\frac{1980-1960}{1945-1940} = 4$ can be interpreted in a meaningful way as the first period being four times as long as the second one. However, the ratio $\frac{1980}{990} = 2$ does not have a meaning; 1980 is not twice as old as 990, hence, the name interval-scale. In addition to statistics at the lower measurement levels, means and standard deviations can be used meaningfully for interval-scaled data.

The next level is ratio-scaled data with examples length and weight. A ratio-scaled variable has no natural unit. Length can be expressed in meters or centimeters, but it has a natural zero. Hence, the permissible transformation is $x' = \alpha x$ $(\alpha > 0)$. In addition to the lower measurement levels, also coefficients of variation can be used meaningfully for ratio-scaled data.

The highest degree of measurability is absolute scale data, e.g., count data. Such data has a natural zero and a natural unit, and the only permissible transformation is the identity. Apart from the measurement levels mentioned above, there are still other types of more exotic scales.[10]

When considering the simple questions, it is clear that metabolomics measurements can have different measurement

https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621

scales. It is certainly not always the case that metabolomics measurements are measured on a ratio-scale. If simple questions 1 and 2 are answered affirmative, then the data is at least ordinal-scaled. If simple questions 3 and 4 are answered affirmative, then the data is at least interval-scaled; and if simple questions 5 and 6 are answered affirmative, then the data is ratio-scaled. This will be explained in the next section.

LEVELS OF METABOLOMICS MEASUREMENTS

Level 0 Measurements: Raw Numbers. Given the knowledge explained above regarding different aspects of comparability, we now turn to the simple questions. The most basic measurement readouts of an LC−MS measurement of blood-lipids are shown in Figure 4a. This is simply a list of raw intensities measured per sample in an LC−MS run arranged in an rt.mz format and will be called level 0 numbers.

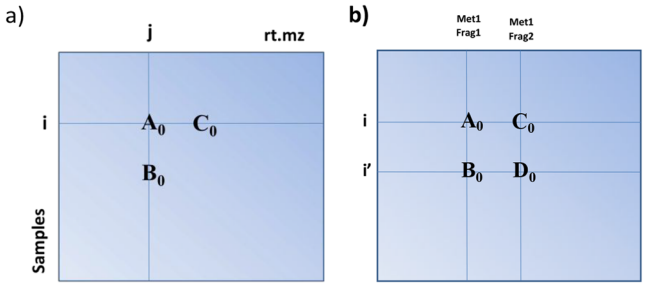


Figure 4. Level 0 of LC−MS measurements: (a) rt.mz is a specific combination of retention time and m/z ratio, i is an index for sample, j is an index for column; (b) Met1 means metabolite 1, Frag1 is fragment 1, and Frag2 is fragment 2 of the same metabolite 1.

We can now start by answering the first simple question. Suppose that $A_0 > C_0$: does that have a meaning? There are two cases to consider. Case a, where the numbers pertain to fragments of different metabolites (we do not consider trivial cases of spurious signals due to noise). For this case, the answer is that $A_0 > C_0$ has no meaning since the response factors of both metabolites are different and at this point unknown (see Supporting Information, Calibration Models). Hence, these numbers do not reflect (relative) concentrations within the system. Case b is shown in Figure 4b and pertains to intensities of different fragments (this does not hold for adducts; their ratios can depend on the concentrations). of the same metabolite (and, thus, at the same rt). In that case, the ratio A_0/C_0 may have meaning since it refers to the same metabolite. In fact, such a ratio should also hold for the same fragments of that metabolite in other rows, thus $A_0/C_0 = B_0/D_0$ (assuming alignment of rt.mz values). Comparing A_0 with B_0 in case a, we run into a lack-of-invariance problem since the rt.mz values are not aligned. Even when alignment would not be a problem, we still have only row-conditional numbers since there may be batch and sample workup differences between the samples.

Level 1 Measurements: Alignment, QC, and IS-Corrected. One of the first steps being done after acquiring the raw data is alignment of the chromatograms, global-IS correction and QC correction of the data (see the Supporting Information, Internal Standards). Alignment is needed to combat the lack-of-invariance problem by assuring that the same feature is now represented in a single column so that each column represents the same compound. Global IS correction is used to reduce sample workup and injection volume errors. The QC correction step is needed to reduce the within and between measurement batch drift of the instruments.[15]

After this data cleaning, we arrive at Level 1 measurements (Figure 5). The columns now represent features and have the

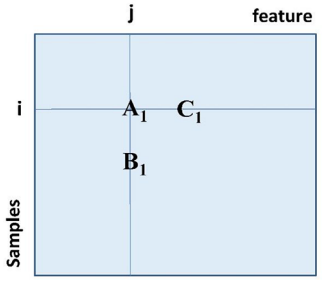


Figure 5. Level 1 of LC−MS measurements after Global-IS and QC correction.

same meaning across each column. A feature can represent one individual lipid molecule but can be also due to a combination of two or more lipid molecules which are isomers but cannot be differentiated with the mass spectrometric method. An example is phosphatidylcholine PC (22:1/18:O) where without MS/MS, the position of the unsaturated fatty acid cannot be determined, and the position of the double bond requires even further advanced methods.

For comparing A_1 and C_1, the same argument goes as for the Level 0 measurements. Comparing A_1 with B_1 is now meaningful since they pertain to the same feature and the numbers are column-conditional due to the IS and QC steps. Still, A_1 and B_1 are measured intensities and not directly interpretable as concentrations. In general, a calibration model has four regions: (i) a below limit of detection region, (ii) a linear region, (iii) a concave region (flattening), and (iv) a saturation region (see the Supporting Information, Calibration Models). If A_1 and B_1 are both in region ii, then their ratio can be interpreted as a ratio of concentrations. Hence, the numbers are ratio-scaled. If they are both in the concave region, then a ratio is not meaningful anymore but if $A_1 > B_1$, then it can still be concluded that the concentration at measurement A_1 is larger than the concentration at measurement B_1. Hence, the numbers are ordinal-scaled. (actually, a bit more than ordinal-scaled since the calibration model has a specific shape.) If one or both of A and B are in the saturated region iv, then a comparison is meaningless. Summarizing, the conclusion about comparability in this case depends crucially on the shape of the calibration model which is unknown at this point.

Until now, we have been discussing numbers. By using instrumental analysis theory into the transition from level 0 to level 1, we have arrived at data because the numbers in level 0 have become a certain meaning. In short: data = numbers + meaning.

Level 2 Measurements: Group IS-Corrected. It is also possible to have internal standards for a group of lipids, e.g., separate standards for triglycerides and certain phospholipid classes such as phosphoethanolamines, phosphoethanolserines, and cholesterols. Compared to level 1, the signal of the feature representing one individual lipid or a combination of isomeric lipids is better quantified as a more appropriate IS is used; the IS should be chosen such that the IS is matching the structure

https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621

of the lipid such that effects such as ion suppression is compensated. Hence, at this level, the metabolite or lipid class of a feature has to be identified. The results are called level 2 measurements and shown in Figure 6.

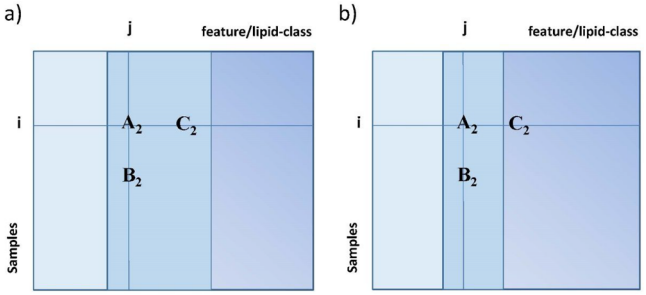


Figure 6. Level 2 of LC−MS measurements after group-IS and QC correction. (Panel a) A_2 and C_2 are in the same lipid class (indicated by shades of blue) and (panel b) A_2 and C_2 are in different lipid classes.

For comparing A_2 and B_2, the conclusions are the same as for the level 1 measurements. In comparing A_2 and C_2, we now have to distinguish between A_2 and C_2 in the same lipid-class or not. When A_2 and C_2 are in the same lipid-class (Figure 6, panel a) and if we can expect similar response factors, then these numbers are comparable. Whether the numbers A_2 and C_2 are ratio- or ordinal-scaled depends again on the region of the calibration models in which A_2 and C_2 are. If the numbers A_2 and C_2 are in different lipid classes (Figure 6, panel b), then we have in principle again level 1 measurements. Note that in the transition from numbers to data, we have not only used instrumental analysis theory but also chemical theory, in particular, theory regarding the behavior during analysis (ionization) and chemical similarity between lipids.

Level 3 Measurements: Concentrations. The highest level of measurements is obtained after having built calibration models for all individual lipids. Obviously, for each rt.mz feature, the structure of the lipid has to be known, which is not a trivial task (but outside of the scope of this paper). At this level, concentrations of a lipid are determined rather than a relative concentration, i.e., a ratio versus an internal standard. This requires that an authentic standard is available, and that thus the lipid is fully identified. An example is the quantification of prostaglandin E2, a bioactive lipid, where absolute concentrations measured in a patient can then be compared to reference values. For this, the most ideal IS is an isotopically labeled prostaglandin E2 eluting at the same retention time and experiencing the same ion suppression. For practical reasons, often calibration models within a class of lipids are constructed using only a limited number of standards. This is possible if the response factors are proven to be the same or a (preferably on theory based) model for the response factor of each lipid is applied[16] (see the Supporting Information, Internal Standards and Calibration Models). These allow for a transformation from intensities to concentrations for all numbers in the matrix of Figure 7. This results in matrix-conditional data, and the data are ratio-scaled.

Level 4 Measurements: Biological Activities. Up until now, the focus has been on concentrations of the lipids. Suppose that the interest is in the lipids as ligands in a biological activity study. It is known that ligands have an

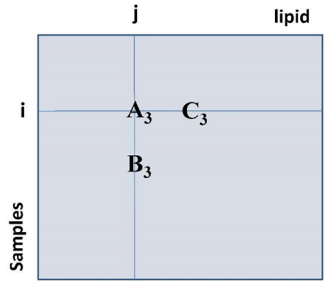


Figure 7. Level 3 of LC−MS measurements after using calibration models.

affinity for a receptor which can be modeled by a (nonlinear) sigmoidal dose−response function which is usually specific for each lipid. An example of such a bioactive lipid is again prostaglandin E2. At level 4 measurements, the bioactive effect, e.g., the pro-inflammatory effect on the vasculature in an in-vitro model, is measured rather than the concentration. From this perspective, the data in level 3 are now suddenly column-conditional since the values A_3 and C_3 have become incomparable due to the differences in dose−response functions. Moreover, because of the sigmoidal relationship, the values A_3 and B_3 are not ratio-scaled anymore but ordinal-scaled (actually, a bit more than ordinal-scaled since the dose−response curves are sigmoidal.) When all dose−response curves are known, then the data could be transformed to biological activities again and become ratio-scaled matrix conditional. This could be called level 4 data, which are tailor-made for a specific purpose.

SYNTHESIS

From the previous presentation, it is clear that there is an interplay between numbers, data, theory, and type of biological question. An attempt to synthesize this is shown in Figure 8.

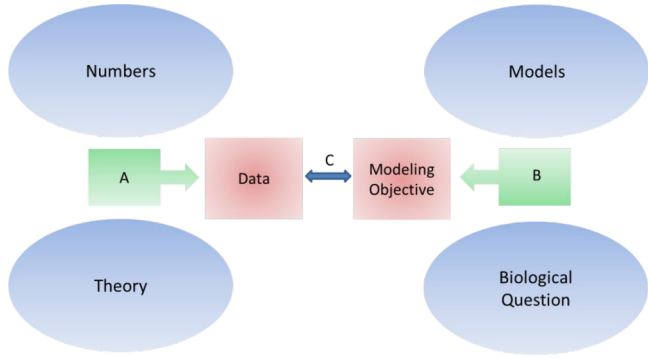


Figure 8. Synthesis of the foregoing. Legend: arrow A represents the transition from numbers to data using chemical- and instrumental analysis theory; arrow B represents the translation from a biological question to a model and modeling objective; and, finally, arrow C asks whether A and B are properly matched.

The blue ellipsoid marked "Numbers" represent the raw numbers coming from an instrument. They do not represent data yet, as explained above. Instrumental analysis and chemical/physical theory should be used to turn these numbers into data (arrow A). These data have then certain properties, conditionality, measurement scale, depending on the original numbers and the theory that is used to turn them into data. This is exemplified above in the different levels of

https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621

metabolomics measurement with increasing efforts to change the properties of the data, e.g, by using internal standards going from level 0 to level 1 and using calibration models going from level 2 to level 3.

The biological questions pertain to certain biological systems, and these questions need to be formalized in a model to be able to confront the question with the data. The term model should be taken in a broad context, e.g., even simple correlations can be considered models. The modeling objective is then formulated in terms of which parameters have to be estimated, which loss-functions to use, which algorithms to use, etc. As an example, if the blood-lipids are measured for a group of controls and patients and if the data are (at least) interval-scaled, then OPLS-DA can be used to find biomarkers.

The crucial part of Figure 8 is arrow C. There should be a match between modeling objectives and properties of the data. Citing the example above, if time-series data of the lipids are available for different subjects and these are not synchronized (or cannot be synchronized), then it does not make sense to use three-way models. If the data are only ordinal scaled, then we cannot fit quantitative systems biology models to the data. If there are discrepancies in arrow C, then there are two routes to take: change the properties of the data or change the modeling objective. For the example of unsynchronized time-series data, we have to switch to simultaneous component analysis (see the section Data Theory) models (thereby possibly also rephrasing the biological question). To fit systems biology models, we have to make calibration models for all lipids and make all data in the concentration form. Obviously, there are many examples of how to solve such discrepancies.

BROADER CONTEXT: CONSIDERATIONS FOR THE FIELD

Repercussion for Metabolomics Data Analysis. The above presented theory has repercussions for metabolomics data analysis. In the subsection Correlations, we will show what it means for correlations as a simple example but similar types of considerations hold for more complicated methods such as PCA and OPLS-DA since these methods use correlations. In the subsection Overview, we will subsequently give an overview.

Correlations. To show how correlations are affected by different comparability properties, we present a small example. Suppose that intensities of three lipids are measured at level 1 (global-IS and QC corrected and aligned). The data for five samples are presented in eq 1:

$$\begin{bmatrix} 10 & 15 & 12 \\ 12 & 10 & 8 \\ 8 & 5 & 4 \\ 10 & . & . \\ 6 & . & . \end{bmatrix} \quad (1)$$

This data is column-conditional and, depending on whether the lipids are measured within the linear range of the calibration models, ordinal or ratio-scaled within a column. The Pearson correlation matrix of this data is

$$\begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \quad (2)$$

assuming that the numbers are ratio-scaled. Suppose now that we have made calibration models for all three lipids and the concentrations are as follows:

$$\begin{bmatrix} 5 & 3 & 6 \\ 6 & 2 & 4 \\ 4 & 1 & 2 \\ 5 & . & . \\ 3 & . & . \end{bmatrix} \quad (3)$$

then this matrix has exactly the same (Pearson) correlations as the one of eq 1 (all intensities were in the linear range of the calibration models). Hence, the column-conditionality of the data does not hamper the use of correlations, and when using correlations, there is no need for calibration models. The reason is that going from intensities to concentrations (assuming that the numbers are in the linear range of the calibration models) are simple linear transformations and correlations are invariant under such linear transformations. If the original intensities were on an ordinal-scale, then similarly Spearman correlations could be used.

Following our example, suppose that we are interested in biological activities and have measured these activities corresponding to the above-mentioned concentrations of prostaglandin E2 and related lipids, and these activities are

$$\begin{bmatrix} 13 & 2 & 3 \\ 15 & 2 & 2 \\ 12 & 1 & 1 \\ 13 & . & . \\ 9 & . & . \end{bmatrix} \quad (4)$$

where lipid two is in the saturation phase of the dose−response curve; lipid one also shows nonlinear behavior, and lipid three is in the linear range. The correlation matrix of these activities is

$$\begin{bmatrix} 1 & 0.76 & 0.33 \\ 0.76 & 1 & 0.87 \\ 0.33 & 0.87 & 1 \end{bmatrix} \quad (5)$$

which is clearly different from eq 2 because of the nonlinearity of the dose−response curves.

Overview. This section discusses the repercussions of the foregoing discussion for metabolomics data analysis. It should not be read as a cookbook about what (not) to do but merely as some remarks about things to consider when performing metabolomics data analysis. Table 2 summarizes the remarks.

The notions of conditionality and measurement scales were explained in the previous sections and summarized in Table 2. As also explained in the foregoing, the data obtained from level 1 and level 2 can be ordinal or ratio- scaled depending on the form of the calibration model and the specific measurement. When we are in the ordinal-scaled regime, then nonmetric methods can be applied such as the Mann−Whitney two-sample tests and nonmetric multidimensional scaling.[17] Also optimal scaling for multivariate analysis is then an option.[18] When we can assume ratio-scaled data (and at least level 2) then the whole (metric) machinery of PCA, PLS, and OPLS-DA is at our disposal. When the data is in the ordinal-scaled regime and still methods such as PCA and OPLS(-DA) are applied, it is unclear at this point whether the results from such

https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621

Table 2. Different Levels of Metabolomics Measurements and Their Properties[a]

level	characteristics	data properties	statistics
level 0	raw numbers	incomparable	some within-row comparisons
level 1	QC-corrected/ aligned	column-conditional	within-column
	global-IS-corrected	ordinal or ratio	nonmetric or metric
level 2	QC-corrected/ aligned	column-conditional	within-column
	group-IS-corrected	within-group matrix-conditional ordinal or ratio	within-group submatrix nonmetric or metric
level 3	concentrations	matrix conditional	within matrix comparisons
		ratio	metric
level 4	tailor made	case specific	case specific

[a]For explanation, see the text.

an analysis are (in)valid: as mentioned earlier the data is also a bit more than ordinal-scaled.

FAIR Data. Recently, the life sciences and especially the omics field starts to agree that data should be FAIR (findable, accesible, interoperable, reusable). This allows to reuse data or to combine data from different sources. However, so far often not much information is provided about the quality and theory of the data: how secure is the identification of a metabolite or lipid? How quantitative are the data: is the data for the metabolites ratio-scaled or ordinal-scaled? If FAIR data is not provided with the proper measurement information and theory (i.e., meta-data), they are actually more numbers than data (see Figure 8).

Data Fusion. A field of growing interest is data fusion and, specifically, fusion of metabolomics data with other types of omics data. The issues of scale-type and comparability (in general, data characteristics) also play a dominant role in this field but until to now have received little attention. An obvious question to ask is whether two data sets can be compared, likewise as comparing two columns in a matrix of measurements as explained above. When different types of omics measurements are performed on the same set of samples, then such questions arise when the two data sets are going to be fused.

Differences in scale-type between two omics data sets also often occur, e.g., when fusing metabolomics with mutation data which are intrinsically binary. Several methods exists for fusing such types of data,[19−22] but comparability issues as explained above have received little attention.

NMR. For NMR, the situation is different than for MS-based metabolomics and we will briefly explain the levels 0−4 for NMR. At level 0, the raw NMR data is considered. These are row-conditional since values in the same column cannot be compared without preprocessing for two reasons. First, the NMR-spectra may not be aligned so that there is lack-of-invariance and, second, even if the spectra are aligned there may still be dilution effects (e.g., in urine spectra) hampering between sample comparisons. Within a row (that is, within the same spectrum) the numbers are comparable because they all pertain to counts of hydrogen atoms.

If all preprocessing has been done (aligning, calibration (e.g., ERETIC signal), and normalization) then we arrive at levels 1−2. The data are now row- and column conditional, hence, matrix conditional. In essence, the data pertains to counts of hydrogen atoms and are not concentrations yet. To arrive at concentrations, the peaks have to be identified, quantified, and calibrated thus thereby arriving at level 3 which can be done by current software such as mNOVA and Chenomx. These concentrations are ratio-scaled and matrix conditional. Also in this case, if interest shifts to biological activities, then the same conclusions (for level 4) hold as for the case of MS-based measurements.

Other Omics Measurements. We do not give a full treatment here, but much of the theory explained above also holds for other types of omics measurements. MS-based proteomics is a clear example, but similar simple questions treated in this paper can also be asked about, e.g., RNAseq data as collected in gene-expression measurements or in microbiome research. We invite researchers in those areas to consider these simple questions too!

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.9b05613.

Short explanation of the methods cited, formal treatment of measurement scales, calibration models, and internal standards (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Age K. Smilde − *Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, 1098 XH Amsterdam, The Netherlands;* orcid.org/0000-0002-3052-4644; Email: a.k.smilde@uva.nl

Author

Thomas Hankemeier − *Analytical Biosciences, LACDR, Leiden University, 2333 CC Leiden, The Netherlands;* orcid.org/0000-0001-7871-2073

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.9b05613

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.K.S. thanks Iven van Mechelen (KU Leuven, Belgium) for useful discussions.

■ REFERENCES

(1) Koek, M.; Jellema, R.; van der Greef, J.; Tas, A.; Hankemeier, T. *Metabolomics* **2011**, *7*, 307−328.

(2) Coombs, C. H. *A Theory of Data*; John Wiley & Sons: New York, 1964.

(3) Van Mechelen, I.; Smilde, A. K. *Chemom. Intell. Lab. Syst.* **2011**, *106*, 2−11.

(4) Zelikovic, I.; Chesney, R. W. *Kidney Int.* **1989**, *36*, 351−359.

(5) Christin, C.; Hoefsloot, H. C. J.; Smilde, A. K.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *J. Proteome Res.* **2010**, *9*, 1483−1495.

(6) Smilde, A. K.; Bro, R.; Geladi, P. *Multiway Analysis: Applications in the Chemical Sciences*; John Wiley & Sons: New York, 2004.

(7) Timmerman, J.; Kiers, H. *Psychometrika* **2003**, *68*, 105−121.

(8) Wold, S.; Kettaneh, N.; Friden, H.; Holmberg, A. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 331−340.

(9) Stevens, S. *Science* **1946**, *103*, 677−680.

(10) Krantz, D.; Luce, R.; Suppes, P.; Tversky, A. *Foundations of Measurement*, Vol. *I*; Dover, 1971.
(11) Roberts, F. Measurement theory. In *Encyclopedia of Mathematics and Its Applications*; Rota, G., Ed.; Cambridge University Press, 1985; Vol. *7*.
(12) Narens, L.; Luce, R. D. *Psychological Bulletin* **1986**, *99*, 166−180.
(13) Luce, R. D.; Narens, L. *Science* **1987**, *236*, 1527−1532.
(14) Hand, D. *Measurement Theory and Practice: The World Through Quantification*; John Wiley & Sons, 2004.
(15) Van der Kloet, F.; Bobeldijk, I.; Verheij, E.; Jellema, R. *J. Proteome Res.* **2009**, *8*, 5132−5141.
(16) Wang, M.; Wang, C. Y.; Han, X. L. *Mass Spectrom. Rev.* **2017**, *36*, 693−714.
(17) Borg, I.; Groenen, P. *Modern Multidimensional Scaling*; Springer, 2005.
(18) Gifi, A. *Nonlinear Multivariate Analysis*; John Wiley & Sons, 1990.
(19) Mo, Q.; Wang, S.; Seshan, V. E.; Olshen, A. B.; Schultz, N.; Sander, C.; Powers, R. S.; Ladanyi, M.; Shen, R. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 4245−4250.
(20) Song, Y.; Westerhuis, J. A.; Aben, N.; Wessels, L. F. A.; Groenen, P. J. F.; Smilde, A. K. *arXiv* **2018**, 1807.04982.
(21) Anderson-Bergman, C.; Kolda, T.; Kincher-Winoto, K. *arXiv* **2018**, 1808.07510.
(22) Smilde, A. K.; Song, Y.; Westerhuis, J. A.; Kiers, H. A. L.; Aben, N.; Wessels, L. F. A. *J. Chemom.* **2020**, e3200.

https://dx.doi.org/10.1021/acs.analchem.9b05613
Anal. Chem. 2020, 92, 13614−13621