IsoSolve: an integrative framework to improve isotopic coverage and consolidate isotopic measurements by MS and/or NMR

Pierre Millard[1,2,‡,*], Sergueï Sokol[1,2,‡,*], Michael Kohlstedt[3], Christoph Wittmann[3], Fabien Létisse[1,4], Guy Lippens[1], Jean-Charles Portais[1,2,4,5]

[1]TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France.

[2]MetaboHUB-MetaToul, National infrastructure of metabolomics and fluxomics, Toulouse, France.

[3]Institute of Systems Biotechnology, Saarland University, Saarbrücken, Germany.

[4]Université Toulouse III - Paul Sabatier, Toulouse, France.

[5]RESTORE, Université de Toulouse, INSERM U1031, CNRS 5070, Université Toulouse III - Paul Sabatier, EFS, Toulouse, France.

ABSTRACT: Stable-isotope labeling experiments are widely used to investigate the topology and functioning of metabolic networks. Label incorporation into metabolites can be quantified using a broad range of mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy methods, but in general, no single approach can completely cover isotopic space, even for small metabolites. The number of quantifiable isotopic species could be increased, and the coverage of isotopic space improved, by integrating measurements obtained by different methods; however, this approach has remained largely unexplored because no framework able to deal with partial, heterogeneous isotopic measurements has yet been developed. Here, we present a generic computational framework based on symbolic calculus that can integrate any isotopic dataset by connecting measurements to the chemical structure of the molecules. As a test case, we apply this framework to isotopic analyses of amino acids, which are ubiquitous to life, central to many biological questions, and can be analyzed by a broad range of MS and NMR methods. We demonstrate how this integrative framework helps to i) clarify and improve the coverage of isotopic space, ii) evaluate the complementarity and redundancy of different techniques, iii) consolidate isotopic datasets, iv) design experiments, and v) guide future analytical developments. This framework, which can be applied to any labeled element, isotopic tracer, metabolite, and analytical platform, has been implemented in IsoSolve (available at https://github.com/MetaSys-LISBP/IsoSolve and https://pypi.org/project/IsoSolve), an open source software that can be readily integrated into data analysis pipelines.

Stable-isotope labeling experiments are widely used to investigate metabolic networks in the fields of systems biology[1-2], biotechnology[3-4] and biomedical research[5-6]. The most effective approach is to combine ^{13}C-labeling strategies with a detailed analysis of isotope incorporation into metabolites, as measured by mass spectrometry (MS) and/or nuclear magnetic resonance (NMR) spectroscopy[7]. MS provides global isotopic information by quantifying the proportions of molecules with different numbers of tracer isotopes (isotopologue distributions)[8-9], while NMR provides positional information on tracer incorporation at specific positions in the molecules (isotopomer distributions)[10-13] by exploiting the ^{1}H and ^{13}C nuclei via non-decoupled experiments – such as homonuclear ^{1}H-^{1}H-TOCSY and heteronuclear ^{1}H-^{13}C-HSQC experiments.

Each separate NMR and MS method provides partial isotopic information by quantifying specific (sets of) isotopic species. MS(/MS) is used to quantify isotopologue distributions of complete molecules or fragments[8-9, 14], where each carbon isotopologue contains several isotopic species that cannot be distinguished since they all have the same mass[9]. NMR is similarly limited in that it only quantifies a subset of isotopic species since positional information is in general limited to a small part of the carbon skeleton (typically from 1 to 3 carbon atoms depending on the experiment). Recently, ^{15}N- and pure-shift-NMR experiments have successfully been applied to access long-range heteronuclear coupling constants, thereby increasing the number of isotopic species that can be quantified[11, 13]. Nevertheless, none of the available methods provides complete coverage of isotopic space, even for small metabolites.

Integrating measurements from different approaches should expand the range of quantifiable isotopic species[7], thus improving the coverage of isotopic space. This is exploited in ^{13}C-fluxomics studies, where different datasets are frequently integrated using isotopic models of metabolic networks to improve flux quantification[4, 15-16]. However, a major drawback of model-based integrative approaches is their strong dependence on the assumptions and simplifications of the model (e.g. the topology of the metabolic network as defined in the model) and the fact that labeling has to be quantified in several metabolites. As an alternative approach, intuitive reasoning has proven useful in improving the coverage of isotopic space by defining relationships between isotopic measurements directly at the level of the molecule. This has been demonstrated for the combination of two NMR experiments, ZQF-TOCSY and HSQC, which allowed absolute quantification of 4 of the 8 isotopomers of a block of three carbon atoms[10]. A few relationships have also been established between specific MS and NMR datasets[17-18], but the heterogeneity of the isotopic information obtained by MS and NMR makes integrating measurements from these two platforms difficult. Overall, the lack of a generic integrative framework has meant that the potential expansion of isotopic coverage that could be

achieved by combining independent MS(/MS) and/or NMR datasets has remained largely untapped.

In this article, we present a computational framework that can be used to integrate any type of isotopic data. We demonstrate how this framework allows isotopic coverage to be clarified and improved, thereby consolidating isotopic measurements. As a test case, we apply this framework to isotopic analysis of amino acids, which are ubiquitous to life, abundant, and central to many biological questions. They can be analyzed by a broad range of MS and NMR methods so the dataset considered here is representative of the wide range of measurements these analytical platforms can provide.

EXPERIMENTAL SECTION

^{13}C-labeled standard of amino acids. The reference material we used to evaluate the proposed methodology was a biologically produced sample containing ^{13}C-labeled amino acids with a controlled and predictable isotopic composition: the isotopic species of each amino acid are forced to all be present at the same concentration. The nature, production, and qualification of this standard sample have been described in detail previously[9, 19].

Isotopic analyses. We analyzed the ^{13}C-labeled reference material by GC-MS[20], and by HSQC[15] and ZQF-TOCSY[15] NMR experiments, as detailed in the corresponding publications. We also gathered additional (HNCA, HACO-DIPSY, LC-MS) datasets for this reference material from the literature[11, 19].

Implementation of the computational framework. The framework developed in this work was implemented as a Python 3 module named IsoSolve. This can be used both as a command-line tool and as a module imported into Python scripts. The intuitive data input is based on tab separated value (tsv) files. The first (mandatory) input file describes the relationships between the measurements and isotopomers as presented in the Results section. The formalism is universal and can be used for all existing and future types of measurement. The second (optional) input file contains the numerical values of the measurements and their associated standard deviations. If provided, these numerical data are used to consolidate the measurements by solving a non-linear least square problem. The symbolic formulas obtained can be verified by assigning randomly drawn values to isotopomers (and thus to the corresponding measurements) and comparing the randomly drawn and calculated values. Explicit results and details of the calculation process can be consulted in a user-friendly HTML document and/or as python variables for later programming use. IsoSolve also generates isotopically-enriched InChIs for all isotopic species (https://github.com/MSI-Metabolomics-Standards-Initiative/inchi-isotopologue-extension), facilitating its integration into standardized data analysis pipelines. IsoSolve is freely available under open source license (GPL v2) at https://pypi.org/project/IsoSolve. A Jupyter notebook (https://jupyter.org) is also provided at https://github.com/MetaSys-LISBP/IsoSolve_notebook as an introduction to programming applications of the software. This notebook contains the code used to perform all the calculations in this study and generate all the equations and Figures 3-6.

RESULTS AND DISCUSSION

General principle. The essence of the proposed framework lies in the way fundamental relationships between measurements and the underlying isotopic species are exploited. As a rule, isotopic measurements from any method can be expressed as the relative abundance of a (set of) isotopic species in a larger set of species. These relationships can be expressed as a system of equations linking independent measurements through isotopic space. We formalize this principle and illustrate how it can be exploited to integrate any type of measurement using the example of alanine, which is routinely analyzed by ^{13}C-NMR, ^{1}H-NMR and MS methods.

^{13}C-NMR experiments provide information on positional isotopomers through J_{CC} coupling patterns, i.e. on the isotopic content of the carbons bonded to the (labeled and detected) carbon. The $J_{C\alpha\text{-}CO}$ and $J_{C\alpha\text{-}C\beta}$ coupling constants of alanine are typically resolved, so the ^{13}C-NMR signal of the C_α atom has four components (*a*, *b*, *c*, *d*), which correspond to 4 individual isotopic species (010, 110, 011 and 111, where 0 and 1 refer to ^{12}C and ^{13}C, respectively, and where the first digit corresponds to the CO group, the second to C_α and the third to C_β). Their abundance is measured relative to the total amount of isotopic species that contribute to these signals, i.e. all species with a labeled C_α atom. C_α ^{13}C-NMR signals can thus be expressed as:

$$a = \frac{010}{010 + 110 + 011 + 111} \quad (1)$$

$$b = \frac{110}{010 + 110 + 011 + 111} \quad (2)$$

$$c = \frac{011}{010 + 110 + 011 + 111} \quad (3)$$

$$d = \frac{111}{010 + 110 + 011 + 111} \quad (4)$$

It should be stressed that this definition ensures that the measurements always add up to 1, a fact that is subsequently used to simplify formulas, as detailed below. In practice, this means that after measuring integrated intensities in arbitrary units, the measurements have to be normalized to their sum. This can always be done provided at least one signal is quantified.

^{1}H-NMR experiments provide information on specific enrichments via J_{CH} coupling patterns, i.e. on the proportion of ^{12}C (*e*) and ^{13}C (*f*) isotopes in the carbon bonded to the analyzed proton. The signal of the H_α proton can thus be expressed as:

$$e = \frac{000 + 100 + 001 + 101}{000 + 001 + 010 + 100 + 101 + 110 + 011 + 111} \quad (5)$$

$$f = \frac{010 + 110 + 011 + 111}{000 + 001 + 010 + 100 + 101 + 110 + 011 + 111} \quad (6)$$

Finally, the total abundance of the set of isotopic species is set by convention to unity, yielding an additional equation:

$$000 + 001 + 010 + 100 + 101 + 110 + 011 + 111 = 1 \quad (7)$$

Measurements obtained by ^{1}H- and ^{13}C-NMR can be integrated by solving this system of equations (eqs 1-7). Expressing the abundance of all isotopic species as a function of the measurements yields the following solution:

$$000 + 001 + 100 + 101 = e \quad (8)$$
$$010 = f \cdot a \quad (9)$$
$$011 = f \cdot b \quad (10)$$
$$110 = f \cdot c \quad (11)$$
$$111 = f \cdot d \quad (12)$$

This system is undetermined, in that while the summed abundance of the four species (000+001+100+101) can be calculated from the measurements, their individual abundances cannot. Nevertheless, the integration of ^{1}H- and ^{13}C-NMR data yields the absolute abundance of four isotopic species (010, 011, 110 and 111), i.e. their abundance is expressed relative to the total amount of molecule rather than to a subset of species (010+110+011+111). This information is new from an analytical standpoint because it cannot be obtained from individual experiments but only by integrating them, as described previously[10].

In contrast to NMR, MS distinguishes molecular entities in terms of the number of labeled atoms incorporated, i.e. by distinguishing between isotopologues. This information can be obtained for different elementary metabolite units (EMUs), which are defined as moieties comprising distinct subsets of the compound's atoms[17]. The carbon isotopologue distribution (CID) is the vector of isotopologue abundances of a given EMU, where the abundance of each isotopologue is expressed relative to the total amount of molecule. The CID of the EMU containing all the carbon atoms of alanine is represented by a vector $[g, h, i, j]$ and is formally defined by the following equations:

$$g = 000 \quad (13)$$
$$h = 001 + 010 + 100 \quad (14)$$
$$i = 011 + 110 + 101 \quad (15)$$
$$j = 111 \quad (16)$$

MS thus provides four additional equations (eqs 13-16) which can be combined with those derived from the ^{1}H- and ^{13}C-NMR data (eqs 1-7). Solving this extended system of equations yields:

$$000 = g \quad (17)$$
$$001 + 100 = h - a \cdot f \quad (18)$$
$$101 = f \cdot a + e - g - h \quad (19)$$
$$010 = f \cdot a \quad (20)$$
$$011 = f \cdot b \quad (21a)$$
$$011 = f \cdot (a + d - j) \quad (21b)$$
$$110 = f \cdot c \quad (22)$$
$$111 = f \cdot d \quad (23a)$$
$$111 = j \quad (23b)$$

Integrating MS data provides new information, namely the absolute abundances of species 000, 101, and the summed abundance of 001 plus 100. The two latter species cannot be resolved individually but the overall level of under-determination is reduced. There is also redundancy in the system of equations, since for two isotopic species, the abundance can be estimated in two ways: 111 can be quantified from either NMR data (eq 23a) or MS data (eq 23b), and 011 from different combinations of MS and NMR data (eqs 21a and 21b).

This intuitive example with ^{1}H-NMR, ^{13}C-NMR and MS data highlights how symbolic calculations can be used to develop a generic framework for integrating isotopic measurements. The calculations are purely based on the fundamental relationships between measurements and the underlying isotopic species. The proposed framework can integrate measurements from any analytical platforms to identify individual species that can be quantified, as well as the combination of species that cannot be resolved individually. This approach thus clarifies the coverage of isotopic space by transforming platform-dependent measurements into (a set of) isotopic species that can be actually quantified. The proposed framework can also consolidate isotopic datasets by integrating quantitative measurements into a single non-linear least squares (NLS) problem. These different aspects are explained in the following sections.

Mathematical formulation and implementation. Eqs 1-4 can be reorganized into linear isotopomer equations with measurements as parameters. For measurement a, this gives:

$$a \cdot (010 + 110 + 011 + 111) - 010 = 0 \quad (24)$$

And linear equations can be obtained similarly for all the other measurements. Let A be the resulting m by p matrix, x a p-length vector of isotopomers, and b the right hand side vector of length m:

$$A \cdot x = b \quad (25)$$

Note that A and b depend on measurements only. The rows of A are linearly dependent as the measurements are normalized to add up to 1.

Eq 25 can be solved by first reducing A to its echelon form by Gauss-Jordan elimination:

$$A = \begin{pmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1p} \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & a_{rr} & \cdots & a_{rp} \\ \vdots & & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \quad (26)$$

The elements in blue can be non-zero while those in black are all 0. The elements on the main diagonal from a_{11} to a_{rr} are strictly non-zero. The number of linearly independent rows defines the *rank* of the matrix, r. Two situations can arise depending on the datasets considered:

- If $r < p$, the system is under-determined, with p-r free isotopomers. Some of the isotopomers depend only on measurements (we call these *defined* isotopomers), while others also depend on *free* isotopomers. It can also happen that some isotopomers are defined by multiple sets of measurements. This is referred to as measurement redundancy, of which eqs 21a,b and eqs 23a,b are examples;
- If $r = p$, the system is just- or over-determined. All isotopomers can be defined uniquely, or in multiple ways in the case of measurement redundancy.

The next step in solving eq 25 is to back-solve the echelon form from x_r to x_1. During echelon reduction and back-solving, the expressions obtained are simplified at each stage using the fact that the measurements from a given method add up to 1. This property was not included in the standard SymPy[21] module we used to manipulate symbolic expressions so we developed dedicated procedures to solve and simplify these expressions. These can be found in the IsoSolve source code.

Once the vector x is obtained as a function of measurements and possibly free isotopomers, measurement redundancy is assessed by substituting isotopomers into the equations defining the measurements, e.g. substituting the solution of eq 25 into eqs 1-4. If in this process, a definition only contains measurements from methods different from the one considered, it is declared redundant.

Formulas for cumomers[22] (i.e. cumulative isotopomers, which describe sets of isotopomers) and EMUs[17] involving measurements only are obtained in a similar way. Solutions for isotopomers are substituted into the equations defining cumomers and EMUs and simplified. Defined cumomers and EMUs are then those without free isotopomers in their final formulas.

When the system is under-determined, one point of interest is: which combinations of isotopomers are still measurable, i.e. not dependent on free isotopomers. This question is addressed by exhaustively testing isotopomers that depend on free isotopomers to identify combinations that can be expressed without free isotopomers. During this procedure, combinations involving shorter, already identified combinations are ignored, such that only elementary measurable combinations are identified.

Figure 1 illustrates how this algorithm is implemented in IsoSolve (Figure 1).

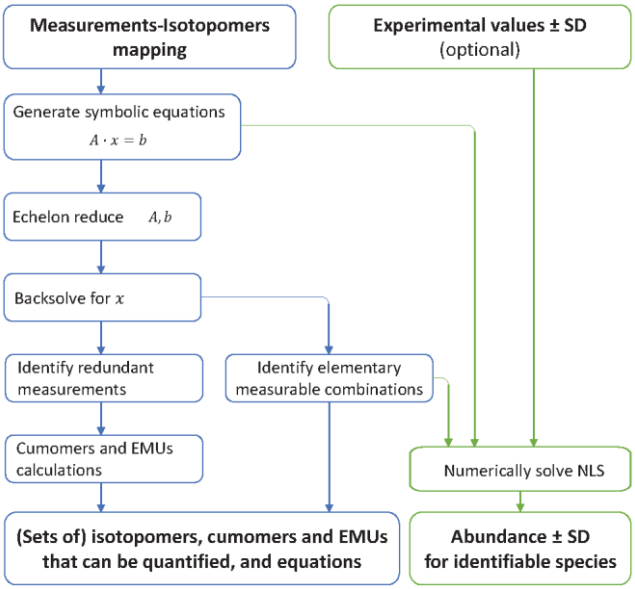


Figure 1. Algorithm implemented in IsoSolve to integrate partial, heterogeneous isotopic measurements. IsoSolve takes as input the relationships between measurements and isotopomers, as defined in Figure 2, to identify the (sets of) isotopomers, cumomers and EMUs that can be quantified individually and produce the corresponding equations (steps framed in blue). When numerical values of measurements are also provided, IsoSolve determines the abundance of all identifiable species (optional steps in green).

A useful feature of IsoSolve that is not directly related to symbolic equation resolution is that experimental data can be consolidated by solving the appropriate NLS problem (Figure 1). The equations in this problem are identical to those defining measurements, such as eqs 1-4. The numerical solution minimizes the cost function *T(x)* defined as the sum of squared residuals. Each residual $u_i(x)$ is calculated as the difference between experimental measurement *i* (y^i_{exp}) and its value calculated from estimated isotopomers ($y^i_{sim}(x)$), normalized by their respective SDs (σ^i_{exp}, provided by the user):

$$min_x\, T(x) = \sum_i \left(u_i(x)\right)^2, \text{ with } u_i(x) = \frac{y^i_{exp} - y^i_{sim}(x)}{\sigma^i_{exp}} \quad (27)$$

Obvious constraints of necessarily non-negative values that add up to 1 are applied to the solution. This NLS problem with inequality and equality constraint is solved using the NLSIC algorithm[23]. A chi-square test is performed to determine if the fit is satisfactory (based on a 95 % confidence threshold). Discrepancies indicate inconsistencies between the different datasets. Finally, IsoSolve estimates the precision of the abundance of each isotopomer, cumomer and EMU by propagating measurement uncertainties. Considering a linearized relationship between small variations in the residual vector Δu and induced variations in the solution vector Δx:

$$J\Delta x = \Delta u \quad (28)$$

Where *J* is the Jacobian matrix of partial derivatives $\partial u / \partial x$, the covariance matrix cov(*x*) is related to a given covariance matrix cov(*u*) as:

$$\mathrm{cov}(x) = J^{-1}\mathrm{cov}(u)(J^{-1})^t \quad (29)$$

Given that *J* is not necessarily invertible, we use a singular value decomposition (SVD) of $J = UD(s)V^t$ where *U* and *V* are orthogonal matrixes and *D(s)* is a diagonal matrix with a vector *s* of strictly positive elements defining the main diagonal. The length of this vector is equal to the rank of *J*, r_J. Since the residuals are scaled by SDs, their covariance matrix is expected to be an identity matrix and the final expression simplifies to:

$$\mathrm{cov}(x) = \frac{r_J}{r_J - 1} VD(s^{-2})V^t \quad (30)$$

Similar to Bessel's correction, the factor $r_J/(r_J - 1)$ ensures that the estimator is not biased. The SDs of *x* are simply the square roots of the elements on the main diagonal of cov(*x*). For sake of brevity, the fact that *x* is constrained to sum to 1 has been omitted from the above description; this constraint is however taken into account in IsoSolve.

Clarifying the isotopic coverage of alanine for individual and combined methods. Combining different analytical methods should improve the coverage of isotopic space. As a first step, we used this workflow to clarify the isotopic information provided by combining a broad range of (NMR and/or MS) methods. Based on the literature[10-11, 13, 19-20, 24-25], we defined a list of eight datasets (D1-D8) that can be obtained for alanine (Figure 2). Even though it only contains three carbon atoms, no single method can completely cover alanine's isotopic space by itself.

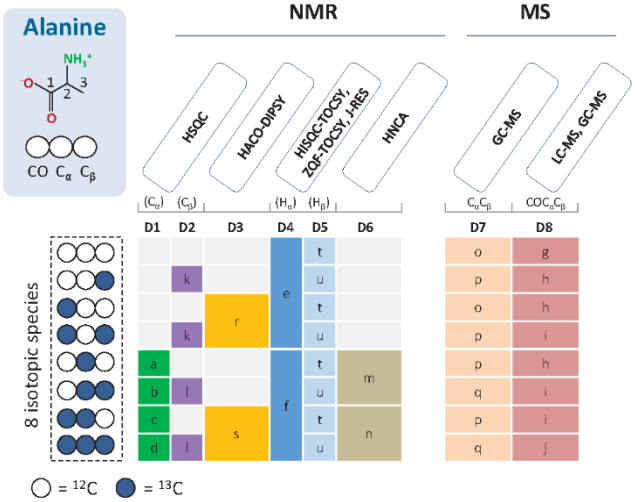


Figure 2. Isotopic measurements on alanine by eight NMR and MS methods. The eight isotopic species of alanine are shown on the left, with white and blue circles representing ^{12}C and ^{13}C atoms, respectively. Methods providing the same information are grouped together (e.g. ZQF-TOCSY and J-RES NMR experiments). For each dataset (D1-D8), each group of measurements is shown by a specific color, and the letters refer to the (sets of) species that are quantified relative to the amount of all species present in the corresponding group.

We evaluated all 255 possible combinations of datasets. The combinations were evaluated based on the following metrics: number of individually quantifiable isotopomers and cumomers, number of EMUs for which all isotopologues can be quantified, number of redundant measurements, and information gainedfrom the proposed integrative framework (i.e. number of additional quantifiable isotopomers).

The results are summarized in Figure 3. In most situations, integrating different datasets improves the coverage of isotopic space, though some combinations do not provide any new information (e.g. D2+D3+D4+D5, combination #127). Importantly, 34 % of the combinations (87/255) provide complete coverage of the isotopic space of alanine (i.e. quantify all its isotopomers, cumomers and EMUs), with different degrees of redundancy (from 0 to 10 redundant measurements). This is only possible if the combined dataset includes both NMR and MS data, highlighting the complementarity of the two techniques. This analysis shows that all the isotopomers of alanine can be quantified with as few as three datasets.

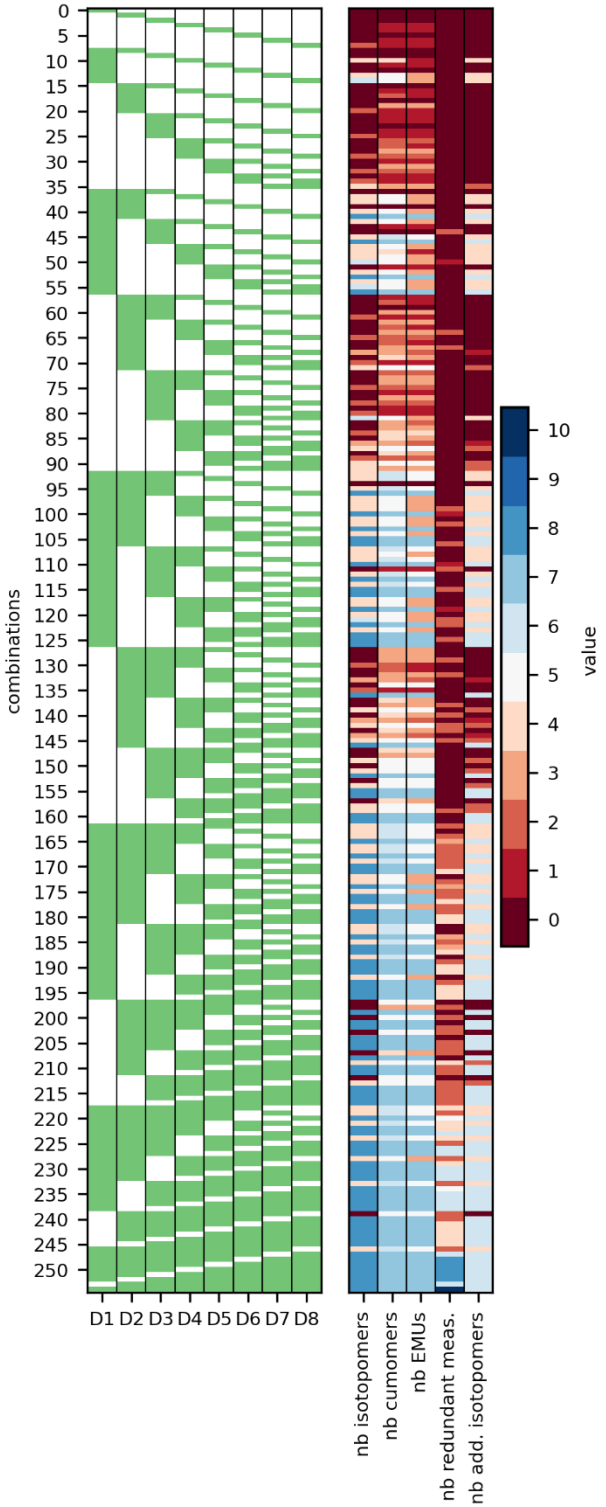


Figure 3. Isotopic information for alanine obtained by integrating different datasets. For each of the 255 possible combinations of datasets (left panel, where each line represents a combination, with included datasets shown in green), the following metrics were calculated (right panel): number of individually quantifiable isotopomers, cumomers and EMUs, number of redundant measurements, and number of additional isotopomers quantifiable only thanks to data integration.

Indeed, combining LC-MS and NMR HSQC data (with C_α and C_β signals, D1+D2+D8, combination #41) leads to the following solution (where the letters refer to the measurements shown in Figure 2):

$$000 = g \quad (31)$$
$$001 = -i - j \cdot \left(1 - (b + d + c \cdot l)/(d \cdot l)\right) \quad (32)$$
$$100 = h + i + j \cdot \left(1 - \left(b + d + l \cdot (a + c)\right)/(d \cdot l)\right) \quad (33)$$
$$101 = i - j \cdot (b + c)/d \quad (34)$$
$$010 = a \cdot j/d \quad (35)$$
$$011 = b \cdot j/d \quad (36)$$
$$110 = c \cdot j/d \quad (37)$$
$$111 = j \quad (38)$$

As well as improving isotopic coverage, this analysis may thus be used to guide experimental design by identifying the best combination of analytical methods and datasets to quantify a given set of isotopic species. As demonstrated here, the complementarity of any technique is readily evaluated, hence providing guidance for future analytical developments.

Isotopic coverage of proteinogenic amino acids. Following the same approach, we used IsoSolve to clarify the current isotopic coverage of proteinogenic amino acids by determining the number of individually quantifiable isotopomers, cumomers and EMUs. Four amino acids are lost during protein hydrolysis (cysteine, tryptophan, glutamine, and asparagine) and cannot be detected. Data integration was thus carried out for the remaining 16 proteinogenic amino acids.

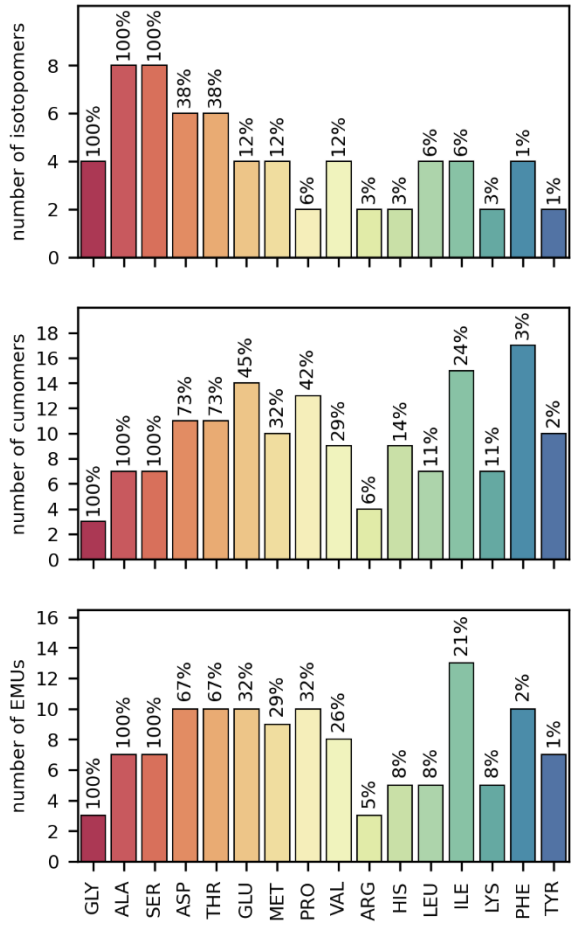


Figure 4. Isotopic coverage of proteinogenic amino acids. Number of isotopomers, cumomers and EMUs that can be quantified individually for each amino acid by integrating available datasets. The respective proportions of isotopic forms that can be quantified are shown above the bars.

Figure 4 highlights the heterogeneity of isotopic coverage for the different amino acids. While complete isotopic coverage is achievable for amino acids containing up to three carbon atoms (glycine, serine and alanine), the coverage progressively decreases as the

number of carbon atoms increases. Isotopomer coverage was 12-31% for C_4-amino acids (aspartate, threonine), 6-12% for C_5-amino acids (glutamate, methionine, proline, valine), 3-6% for C_6-amino acids (arginine, histidine, leucine, isoleucine, lysine), and about 1% for C_9-amino acids (phenylalanine and tyrosine). As expected, the coverages are higher for cumomers (100% for C_2 and C_3, 40-53% for C_4, 23-42% for C_5, 6-24% for C_6, 2-3% for C_9) and for EMUs (100% for C_2 and C_3, 27-40% for C_4, 16-32% for C_5, 5-21% for C_6, 1-2% for C_9) than for isotopomers. Overall, this framework provides a clear audit of the isotopic information that can actually be measured on proteinogenic amino acids.

Data integration and consolidation. To quantitatively evaluate the proposed integrative framework, we produced a reference sample of ^{13}C-labeled proteinogenic amino acids with controlled and predictable labeling patterns in which all isotopic species are present in equal amounts[9].

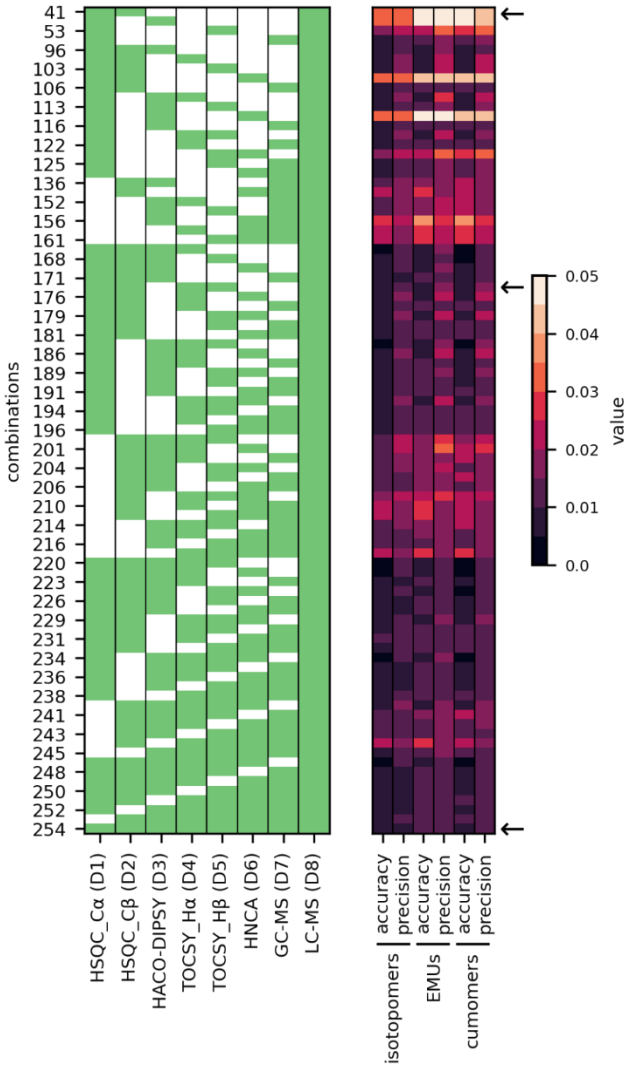


Figure 5. Summary of data integration results for all dataset combinations that completely cover the isotopic space of alanine. The accuracy and precision of isotopomer, cumomer and EMU quantifications (right panel) were determined for each combination of datasets (left panel). Combinations discussed in the text and detailed in Figure 6 are indicated by an arrow.

This sample was analyzed by NMR (ZQF-TOCSY, HSQC, HNCA and HACO-DIPSY experiments) and MS (GC-MS and LC-MS), yielding eight independent datasets (D1-D8 in Figure 2) containing a total of 21 isotopic measurements for alanine (Supporting information S1). For all combinations of datasets identified as providing complete isotopic coverage of alanine (Figure 3), we used IsoSolve to determine the abundance of each isotopomer, cumomer and EMU. The results obtained for each combination were evaluated using two quantitative metrics[19]: accuracy (defined as the mean error and calculated from the differences between theoretical and measured abundances) and precision (defined as the mean standard deviation of the measured abundances).

All the isotopomers, cumomers and EMUs of alanine were indeed quantified for all the combinations considered. The accuracy and precision of the results depended on the datasets included (Figure 5).

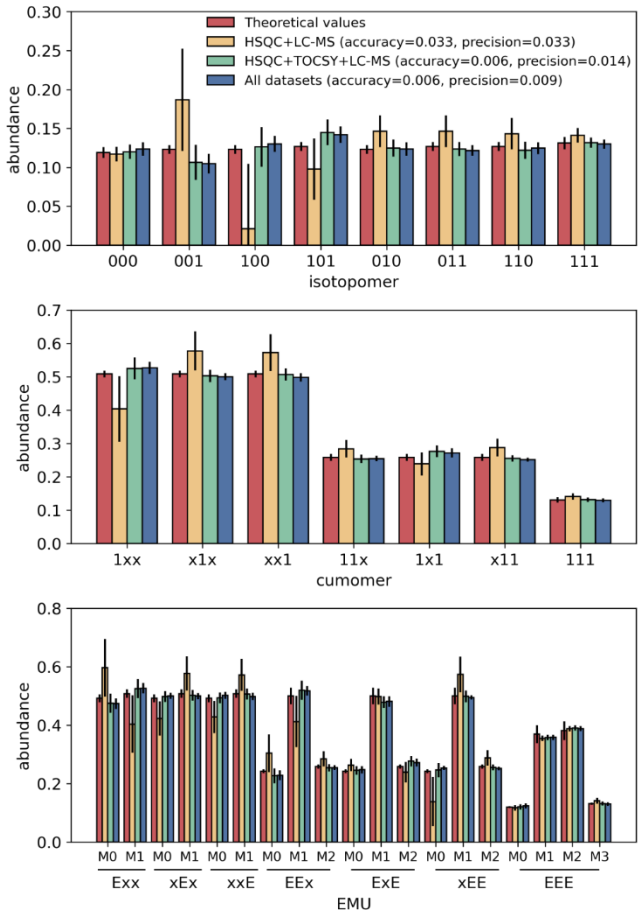


Figure 6. Detailed quantification results for three different combinations of datasets. The abundance of each isotopomer (upper panel), cumomer (middle panel) and EMU (lower panel) was estimated by integrating different datasets (HSQC+LC-MS, orange bars; HSQC+TOCSY+LC-MS, green bars; all datasets, blue bars), and experimental values were compared to the theoretical abundances in the reference sample (red bars). ^{12}C- and ^{13}C-atoms are represented by 0s and 1s, respectively; *x* stands for "0 or 1", and *E*s denote the atoms contained in the corresponding EMU. Error bars correspond to ± one standard deviation.

Regarding isotopomers for example, integrating HSQC and LC-MS datasets (D1+D2+D8, combination #41) led to an accuracy and precision of 0.033. Adding the TOCSY NMR dataset (with H_α and H_β signals, D1+D2+D4+D5+D8, combination #174) improved both metrics (accuracy = 0.006, precision = 0.014) and results were further improved when all datasets were combined (D1-8, combination #254, accuracy = 0.006, precision = 0.009). Similar trends were observed for cumomers and EMUs (Figure 5).

A detailed analysis of the integration results reveals that precision and accuracy also depend on the isotopic species considered (Figure 6). Combining HSQC and LC-MS data is sufficient to reliably quantify 6 of the 8 isotopomers of alanine. Isotopomers 000 (accuracy = 0.002, precision = 0.009) and 011 (accuracy = 0.020, precision = 0.020) are for instance well resolved, but isotopomers 001 (accuracy = -0.060, precision = 0.070) and 100 (accuracy = 0.102, precision = 0.082) remain poorly resolved. Adding the TOCSY dataset significantly improves quantification for the two latter species (001: accuracy = 0.018, precision = 0.022; 100: accuracy = -0.008, precision = 0.025). The most reliable results were obtained by integrating all the datasets (001: accuracy = 0.018, precision = 0.013; 100: accuracy = -0.007, precision = 0.010). Here again, similar trends were observed for cumomers and EMUs (Figure 6).

In combinations with measurement redundancy, the consistency of the different datasets can be evaluated using a chi-square test. For instance, when all datasets were combined (D1-8, combination #254, 10 redundant measurements), the chi-square test confirmed that all the measurements were consistent (p-value = 0.994). When some of these measurements were artificially altered, e.g. *b* changed from 0.2533 to 0.0533 and *c* from 0.2475 to 0.4475, the p-value decreased to 2×10^{-16}, highlighting the inconsistencies between the redundant measurements. This illustrates how the proposed approach can be used to identify biased measurements to be checked before interpretation.

These results confirm that integrating additional datasets improves both the accuracy and the precision of isotopomer quantification, very likely because of the high degree of redundancy (up to 10 redundant measurements) which reduces the impact of experimental noise and the potential biases of individual measurements. Overall, all isotopomers could be reliably quantified using a wide variety of data combinations, with a high accuracy and precision in most situations.

CONCLUSION

The complementarity of different (MS and/or NMR) approaches dedicated to isotopic analyses is often highlighted, but the lack of a generic integrative framework able to deal with heterogeneous, partial isotopic measurements has meant that this has never been evaluated in detail. The proposed framework fills this conceptual gap by allowing any type of isotopic measurements (MS, MS/MS, ^{1}H-NMR, ^{13}C-NMR, ^{15}N-NMR, etc) to be included. The framework is agnostic to the analytical platform, the labeled element, the tracer isotope, and the molecule. It can also be applied to double-labeling approaches (e.g. ^{13}C and ^{15}N). This framework has been implemented as an open source Python program, IsoSolve, which is available as a command-line interface and as a Python library to streamline its integration into existing data analysis pipelines.

Using amino acids as an example application, we have demonstrated that this framework can i) clarify the actual coverage of isotopic space by identifying the (sets of) species that can actually be quantified, ii) improve this coverage by increasing the number of isotopic species that can be quantified individually, iii) evaluate the complementarity and redundancy of different techniques, iv) consolidate isotopic datasets by evaluating their consistency, identifying biased measurements, and reducing the impact of measurement noise, v) support experimental design by identifying the most relevant methods to quantify a given set of isotopic species, and vi) guide future analytical developments.

Our framework connects measurements to the chemical structure of compounds and to their formal representation in isotopic models of metabolism, hence assisting both model-free and model-based data interpretation. This framework may thus support structural investigations (e.g. metabolite identification, spectra annotation, validation of MS/MS fragmentation patterns, development of standardized databases for deposition of isotopic datasets based on isotopically-resolved InChIs) as well as functional investigations of metabolic systems (e.g. experimental design, data consolidation, conversion between isotopic representations in ^{13}C-fluxomics workflows). It should also make isotope labeling experiments more accessible to the wider biological community.

Beyond isotopic studies, this framework may prove equally valuable in other fields dealing with the analysis of combinatorial states of biological entities. This is the case in proteomics for instance, for the analysis of post-translational modifications (e.g. phosphorylation or acetylation), with the ultimate objective of determining the complete distribution of each of the 2^n forms of a protein with n-modification sites. Partial information on these distributions can be obtained by MS and NMR[26], and these datasets can be integrated using IsoSolve following the principles described in this article.

AUTHOR INFORMATION

Corresponding Author

* Telephone: +33(0)5-61-55-93-55. E-mail: pierre.millard@insa-toulouse.fr ; sokol@insa-toulouse.fr.

Author Contributions

‡These authors contributed equally.

Notes

The authors declare no competing financial interest.

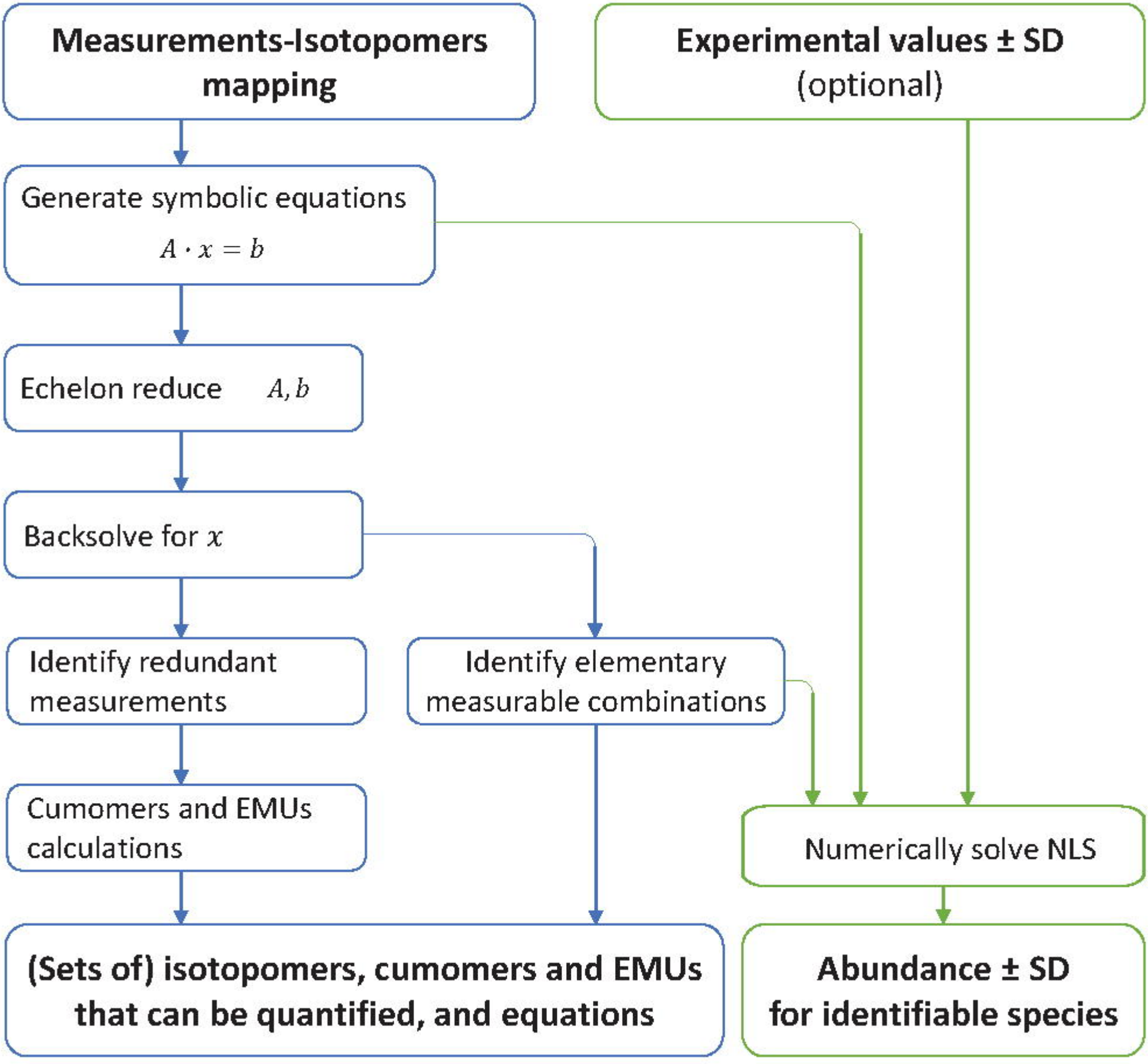
ACKNOWLEDGMENT

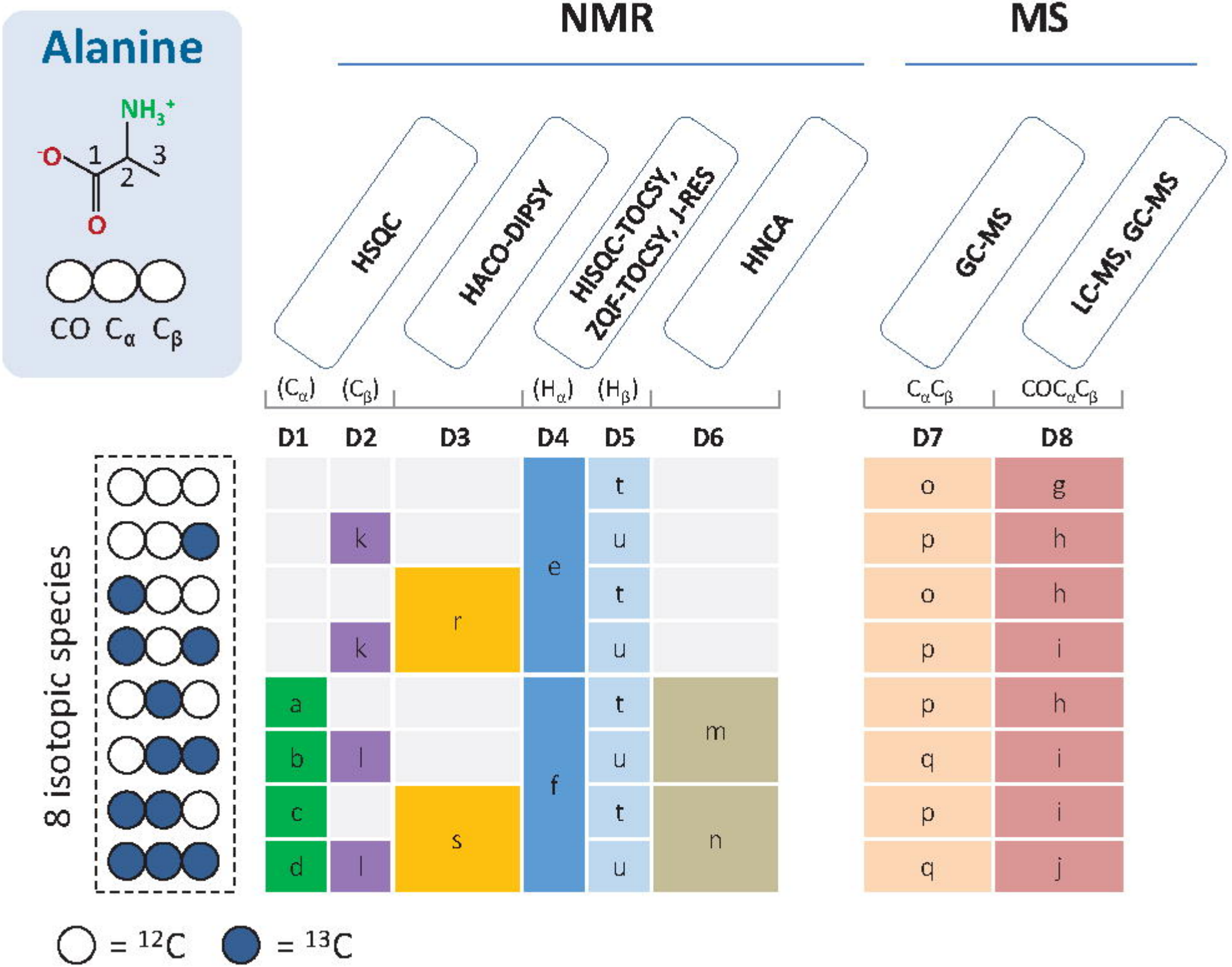
The authors thank MetaboHub-MetaToul (Metabolomics & Fluxomics facilities, Toulouse, France, http://www.metatoul.fr), which is part of the French National Infrastructure for Metabolomics and Fluxomics (www.metabohub.fr), funded by the ANR (MetaboHUB-ANR-11-INBS-0010), for access to NMR and MS facilities. JCP is grateful for funding from INSERM for his temporary full-time researcher position.

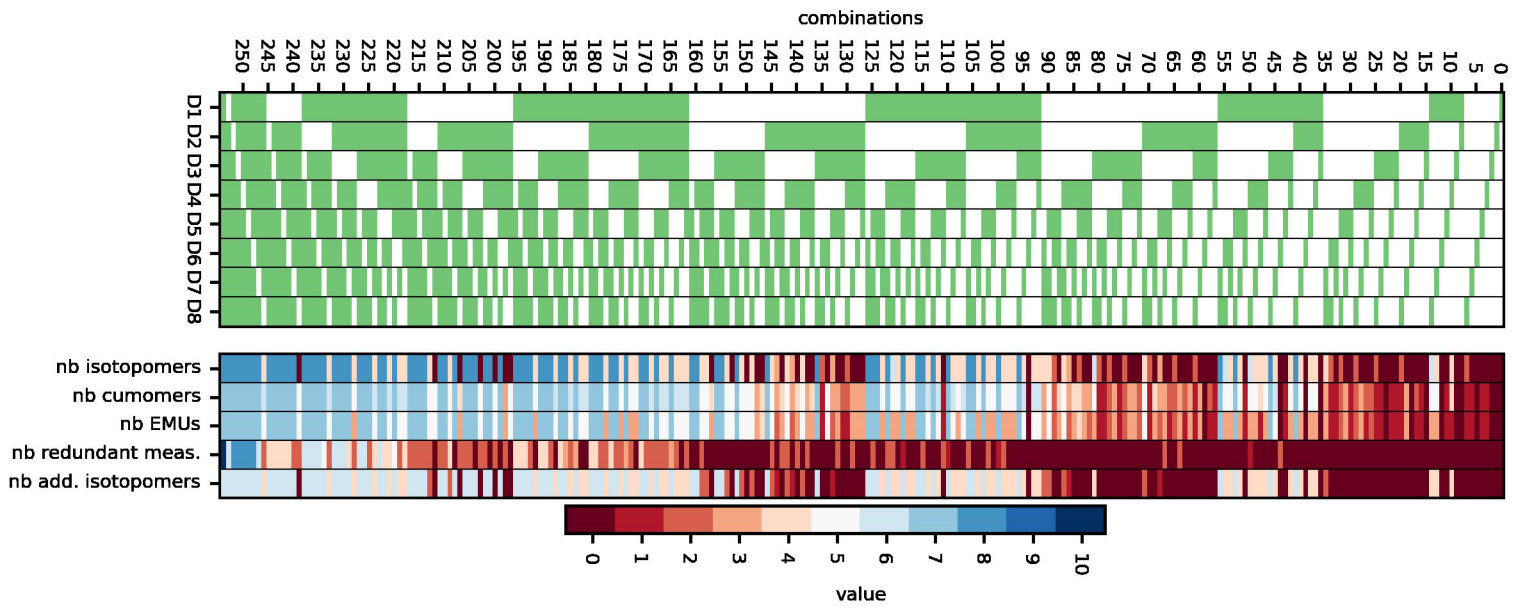
REFERENCES

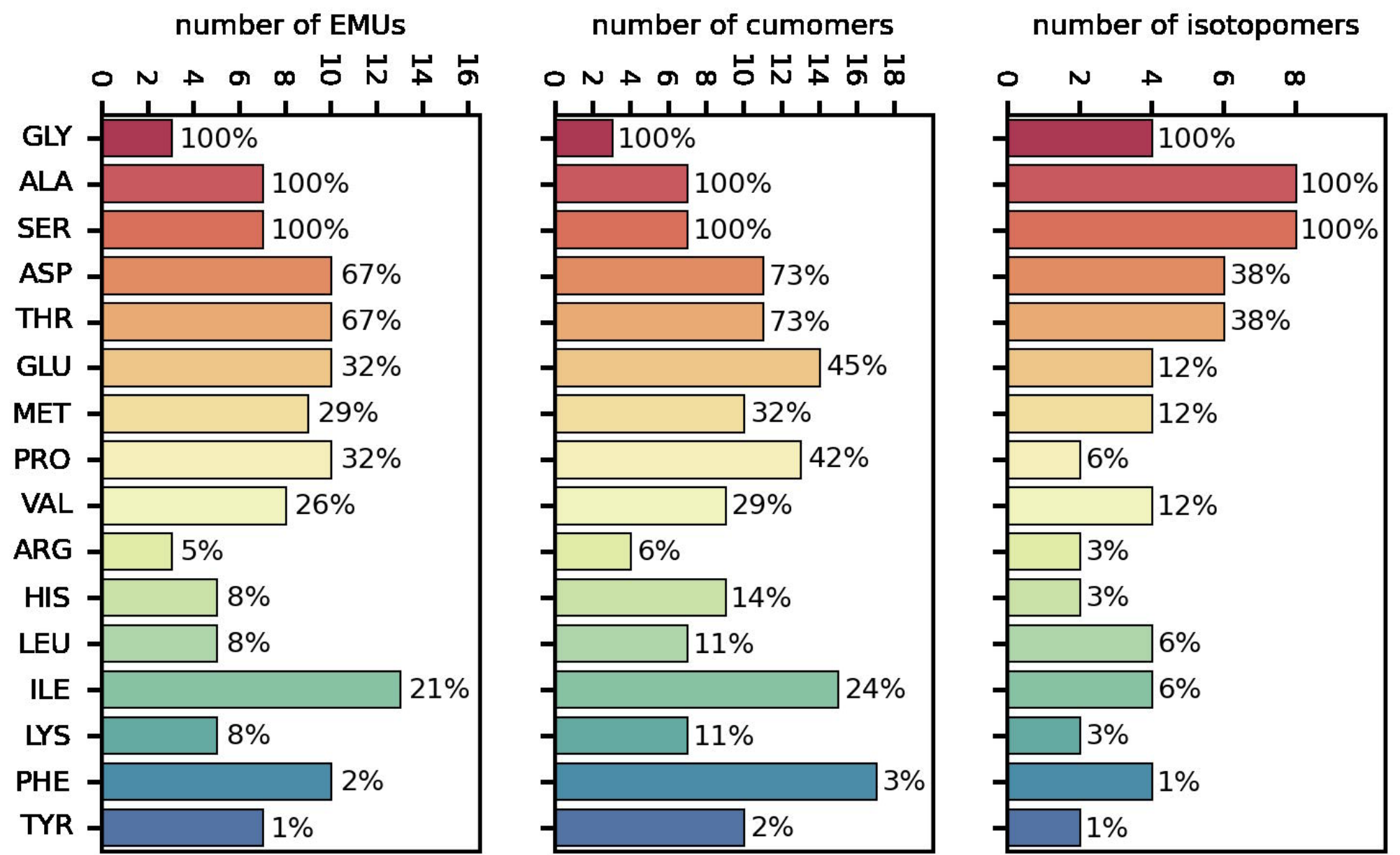
(1) Millard, P.; Enjalbert, B.; Uttenweiler-Joseph, S.; Portais, J. C.; Letisse, F., *BioRxiv preprint* **2020**. doi: 10.1101/2020.08.18.255356.
(2) Christodoulou, D.; Link, H.; Fuhrer, T.; Kochanowski, K.; Gerosa, L.; Sauer, U., *Cell systems* **2018,** *6* (5), 569-578 e7. doi: 10.1016/j.cels.2018.04.009.
(3) Millard, P.; Schmitt, U.; Kiefer, P.; Vorholt, J. A.; Heux, S.; Portais, J. C., *PLoS Comput Biol* **2020,** *16* (4), e1007799. doi: 10.1371/journal.pcbi.1007799.
(4) Heux, S.; Berges, C.; Millard, P.; Portais, J. C.; Letisse, F., *Curr Opin Biotechnol* **2017,** *43*, 104-109. doi: 10.1016/j.copbio.2016.10.010.
(5) Hui, S.; Cowan, A. J.; Zeng, X.; Yang, L.; TeSlaa, T.; Li, X.; Bartman, C.; Zhang, Z.; Jang, C.; Wang, L.; Lu, W.; Rojas, J.; Baur, J.; Rabinowitz, J. D., *BioRxiv preprint* **2020**. doi: 10.1101/2020.03.02.973669.
(6) Han, B.; Wang, L.; Zhang, J.; Wei, M.; Rajani, C.; Wei, R.; Wang, J.; Yang, H.; Carbone, M.; Xie, G.; Zhou, W.; Jia, W., *BioRxiv preprint* **2020**. doi: 10.1101/2020.06.04.132902.
(7) Letertre, M. P. M.; Dervilly, G.; Giraudeau, P., *Anal Chem* **2021,** *93* (1), 500-518. doi: 10.1021/acs.analchem.0c04371.
(8) Wittmann, C., *Microb Cell Fact* **2007,** *6*, 6. doi: 10.1186/1475-2859-6-6.
(9) Millard, P.; Massou, S.; Portais, J. C.; Letisse, F., *Anal Chem* **2014,** *86* (20), 10288-95. doi: 10.1021/ac502490g.
(10) Massou, S.; Nicolas, C.; Letisse, F.; Portais, J. C., *Phytochemistry* **2007,** *68* (16-18), 2330-40. doi: 10.1016/j.phytochem.2007.03.011.
(11) Millard, P.; Cahoreau, E.; Heuillet, M.; Portais, J. C.; Lippens, G., *Anal Chem* **2017,** *89* (3), 2101-2106. doi: 10.1021/acs.analchem.6b04767.
(12) Szyperski, T., *Eur J Biochem* **1995,** *232* (2), 433-48. doi:
(13) Sinnaeve, D.; Dinclaux, M.; Cahoreau, E.; Millard, P.; Portais, J. C.; Letisse, F.; Lippens, G., *Anal Chem* **2018,** *90* (6), 4025-4031. doi: 10.1021/acs.analchem.7b05206.

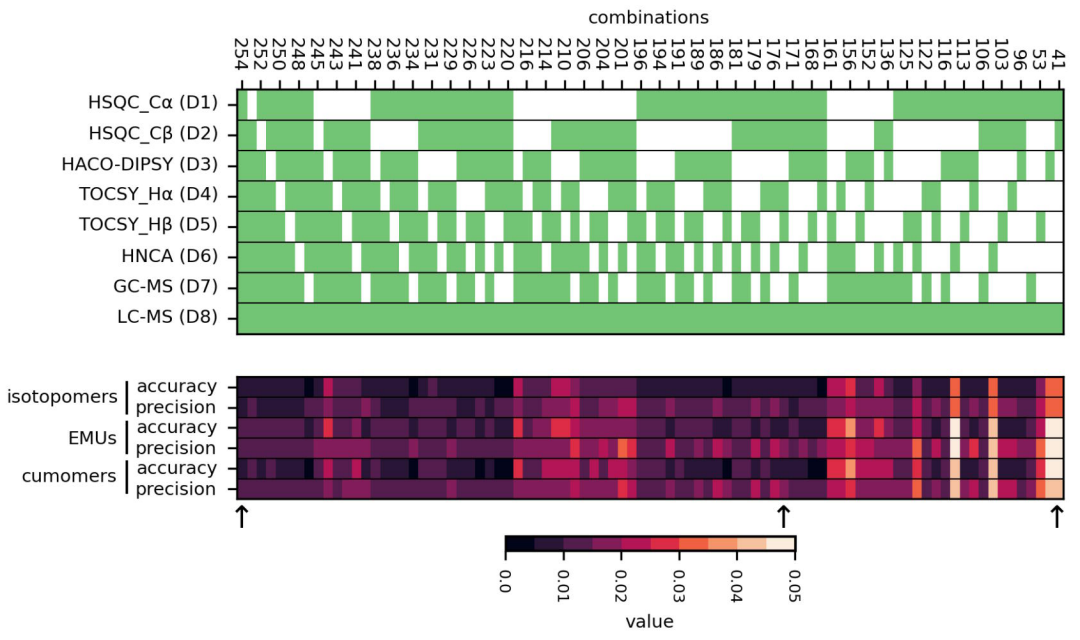
(14) Choi, J.; Antoniewicz, M. R., *Metab Eng* **2011,** *13* (2), 225-33. doi: 10.1016/j.ymben.2010.11.006.
(15) Millard, P.; Massou, S.; Wittmann, C.; Portais, J. C.; Letisse, F., *Anal Biochem* **2014,** *465*, 38-49. doi: 10.1016/j.ab.2014.07.026.
(16) Schwechheimer, S. K.; Becker, J.; Peyriga, L.; Portais, J. C.; Sauer, D.; Muller, R.; Hoff, B.; Haefner, S.; Schroder, H.; Zelder, O.; Wittmann, C., *Metab Eng* **2018,** *47*, 357-373. doi: 10.1016/j.ymben.2018.04.005.
(17) Antoniewicz, M. R.; Kelleher, J. K.; Stephanopoulos, G., *Metab Eng* **2007,** *9* (1), 68-86. doi: 10.1016/j.ymben.2006.09.001.
(18) Deja, S.; Fu, X.; Fletcher, J. A.; Kucejova, B.; Browning, J. D.; Young, J. D.; Burgess, S. C., *Metab Eng* **2020,** *59*, 1-14. doi: 10.1016/j.ymben.2019.12.005.
(19) Heuillet, M.; Bellvert, F.; Cahoreau, E.; Letisse, F.; Millard, P.; Portais, J. C., *Anal Chem* **2018,** *90* (3), 1852-1860. doi: 10.1021/acs.analchem.7b03886.
(20) Kohlstedt, M.; Wittmann, C., *Metab Eng* **2019,** *54*, 35-53. doi: 10.1016/j.ymben.2019.01.008.
(21) Meurer, A.; Smith, C. P.; Paprocki, M.; Čertík, O.; Kirpichev, S. B.; Rocklin, M.; Kumar, A.; Ivanov, S.; Moore, J. K.; Singh, S.; Rathnayake, T.; Vig, S.; Granger, B. E.; Muller, R. P.; Bonazzi, F.; Gupta, H.; Vats, S.; Johansson, F.; Pedregosa, F.; Curry, M. J.; Terrel, A. R.; Roučka, Š.; Saboo, A.; Fernando, I.; Kulal, S.; Cimrman, R.; Scopatz, A., *PeerJ Computer Science* **2017,** *3*, e103. doi: 10.7717/peerj-cs.103.
(22) Wiechert, W.; Mollney, M.; Isermann, N.; Wurzel, M.; de Graaf, A. A., *Biotechnol Bioeng* **1999,** *66* (2), 69-85. doi:
(23) Sokol, S.; Millard, P.; Portais, J. C., *Bioinformatics* **2012,** *28* (5), 687-93. doi: 10.1093/bioinformatics/btr716.
(24) Antoniewicz, M. R.; Kelleher, J. K.; Stephanopoulos, G., *Anal Chem* **2007,** *79* (19), 7554-9. doi: 10.1021/ac0708893.
(25) Cahoreau, E.; Peyriga, L.; Hubert, J.; Bringaud, F.; Massou, S.; Portais, J. C., *Anal Biochem* **2012,** *427* (2), 158-63. doi: 10.1016/j.ab.2012.05.021.
(26) Prabakaran, S.; Everley, R. A.; Landrieu, I.; Wieruszeski, J. M.; Lippens, G.; Steen, H.; Gunawardena, J., *Mol Syst Biol* **2011,** *7*, 482. doi: 10.1038/msb.2011.15.

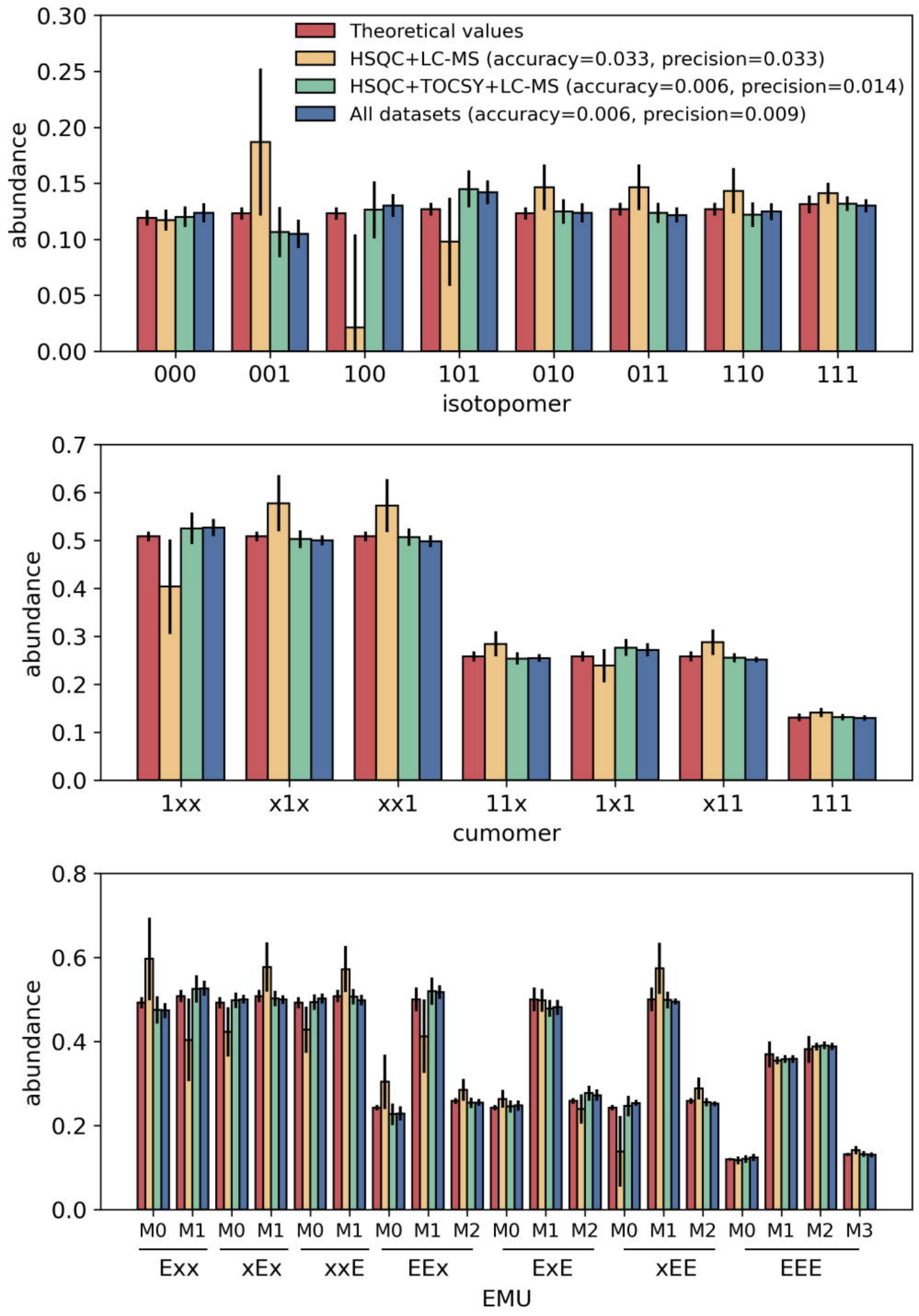

Measurements-Isotopomers mapping
Experimental values ± SD (optional)
Generate symbolic equations
$A \cdot x = b$
Echelon reduce A, b
Backsolve for x
Identify redundant measurements
Identify elementary measurable combinations
Cumomers and EMUs calculations
Numerically solve NLS
(Sets of) isotopomers, cumomers and EMUs that can be quantified, and equations
Abundance ± SD for identifiable species


Alanine
NH3+
O
1
2
3
O
CO Cα Cβ
NMR
MS
HSQC
HACO-DIPSY
HISQC-TOCSY, ZQF-TOCSY, J-RES
HNCA
GC-MS
LC-MS, GC-MS
(Cα) (Cβ) (Hα) (Hβ)
CαCβ COCαCβ
D1 D2 D3 D4 D5 D6 D7 D8
8 isotopic species
a b c d
k k l l
r s
e f
t u t u t u t u
m n
o p o p p q p q
g h h i h i j
○ = 12C ● = 13C


combinations
250 245 240 235 230 225 220 215 210 205 200 195 190 185 180 175 170 165 160 155 150 145 140 135 130 125 120 115 110 105 100 95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
D1 D2 D3 D4 D5 D6 D7 D8
nb isotopomers
nb cumomers
nb EMUs
nb redundant meas.
nb add. isotopomers
0 1 2 3 4 5 6 7 8 9 10
value


number of EMUs
0
2
4
6
8
10
12
14
16
GLY
ALA
SER
ASP
THR
GLU
MET
PRO
VAL
ARG
HIS
LEU
ILE
LYS
PHE
TYR
100%
100%
100%
67%
67%
32%
29%
32%
26%
5%
8%
8%
21%
8%
2%
1%
number of cumomers
0
2
4
6
8
10
12
14
16
18
100%
100%
100%
73%
73%
45%
32%
42%
29%
6%
14%
11%
24%
11%
3%
2%
number of isotopomers
0
2
4
6
8
100%
100%
100%
38%
38%
12%
12%
6%
12%
3%
3%
6%
6%
3%
1%
1%


combinations
254
252
250
248
245
243
241
238
236
234
231
229
226
223
220
216
214
210
206
204
201
196
194
191
189
186
181
179
176
171
168
161
156
152
136
125
122
116
113
106
103
96
53
41
HSQC_Cα (D1)
HSQC_Cβ (D2)
HACO-DIPSY (D3)
TOCSY_Hα (D4)
TOCSY_Hβ (D5)
HNCA (D6)
GC-MS (D7)
LC-MS (D8)
isotopomers
EMUs
cumomers
accuracy
precision
accuracy
precision
accuracy
precision
0.0
0.01
0.02
0.03
0.04
0.05
value


Theoretical values
HSQC+LC-MS (accuracy=0.033, precision=0.033)
HSQC+TOCSY+LC-MS (accuracy=0.006, precision=0.014)
All datasets (accuracy=0.006, precision=0.009)
abundance
0.30
0.25
0.20
0.15
0.10
0.05
0.00
000
001
100
101
010
011
110
111
isotopomer
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1xx
x1x
xx1
11x
1x1
x11
111
cumomer
0.8
0.6
0.4
0.2
0.0
M0 M1 M0 M1 M0 M1 M0 M1 M2 M0 M1 M2 M0 M1 M2 M0 M1 M2 M3
Exx
xEx
xxE
EEx
ExE
xEE
EEE
EMU