*Laboratory Astrophysics: from Observations to Interpretation*
*Proceedings IAU Symposium No. 350, 2019*
*F. Salama & H. Linnartz, eds.*
doi:10.1017/S1743921319008032

# Measured mass to stoichoimetric formula through exhaustive search

**François-Régis Orthous-Daunay[1], Roland Thissen[2] and Véronique Vuitton[1]**

[1]Institut de Planétologie et d'Astrophysique de Grenoble, Univ. Grenoble Alpes, CNRS, CS 40700, 38058 Grenoble Cédex 9, France
email: frod@univ-grenoble-alpes.fr

[2]Laboratoire de Chimie Physique, CNRS, Univ. Paris Sud, Université Paris-Saclay, 91405, Orsay, France

**Abstract.** Electrospray ionisation has revolutionised mass spectrometry. Coupled to high mass resolution, it provides the stoichiometric formula of a lot of molecules in a mixture. The link between the mass spectrometry data and the chemical description relies on an interpretation of the measured masses. We present here the tools and tricks developed to exploit Orbitrap mass spectra. This piece of work focuses on the numerical method to assign a molecular formula to a measured mass. The problem is restrained to the solving of the Diophantine equation where the constant coefficients are stoichiometric groups. Peculiar case of a set of convenient groups is given with the chemical constraints it brings to the problem.

**Keywords.** astrochemistry, molecular processes, methods: data analysis

## 1. Introduction

Planetary and earth science samples like oil (Kozhinov *et al.* (2013)), Titan ana logs (Pernot *et al.* (2010)), cometary analogs (Danger *et al.* (2013)) were extensively studied with Orbitrap instrument (Figure 1).

The Orbitrap is a Fourier Transform mass spectrometer designed by Alexander Makarov (Makarov (2000)) and exploited during the mid 00's (Makarov (2006)). It delivers a mass resolution of $\frac{m}{\Delta m} \geqslant 10^5$ for a measured mass at 400u. The resolution decreases with mass in $\frac{1}{\sqrt{m}}$. The capability of resolving every molecular mass in a mixture depends on its peaks density, which is unknown *a priori*. Thereafter, we assume that ions are produced by electrospray ionisation (Yamashita & Fenn (1984)) and that the mass spectrometry is the one of molecules in the mixture, modulo the addition or subtraction of an integer number of protons. Our method applies to molecular and radical ions alternatively.

The higher the resolution, the higher the number of detected peaks that can be interpreted as molecular masses. Independently, the number of possible different formula that are in an interval around a given mass increases with $m$. In order to cope with the combinatorial explosion of the masses cardinal and the loss of resolution, a series of numerical recipes and assumptions must be made. In the review (Meija (2006)), the author makes a pretty much exhaustive list of mathematical tools that have an implication in the data analysis. A software suite that has been developed at IPAG since 2010 assembles several algorithms in order to produce an integrated environment to handle FT-MS data. This piece of code is called ATTRIBUTOR. The improvement proposed here is a method that computes all the masses that match the measurement at the natural number level. To do

**Table 1.** Masses of interest (u).

| Element | Mass | Groups | Mass |
|---|---|---|---|
| $^{1}$H | 1.007 825 032 23 | $^{12}$C$^{1}$H$_2$ | 14.0156500645 |
| $^{12}$C | 12 | $^{14}$N$^{1}$H | 15.01089903666 |
| $^{14}$N | 14.003 074 004 43 | | |
| $^{16}$O | 15.994 914 619 57 | | |

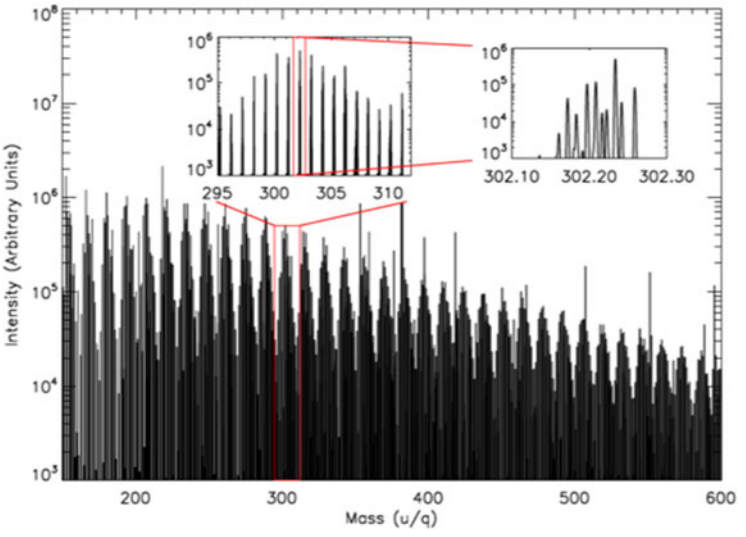


**Figure 1.** Typical Orbitrap mass spectrum of Titan analogs.

so, we demonstrate that a chosen set of stoichiometric groups corresponds to a set of chemical rules when combined with positive integer coefficients. Then we describe our way to efficiently generate the list of Diophantine equation solutions.

## 2. Stoichiometric formula decomposition

The link between a measured mass and the stoichiometric formula in the associated molecule is the linear combination of the mass of the elements. Masses are expressed in the u unit which is the International Union of Pure and Applied Chemistry (IUPAC) recommendation and standard. It sets the mass of $^{12}$C in its fundamental state to 12u. From a relativistic point of view, the binding energy between nucleons is equivalent to inertial mass and therefore could be measured with mass spectrometry. This is observed through the fact that elements have a non integer mass compared to a twelfth of $^{12}$C's mass. Table 1 consists in a crop of the NIST database concerning the mass of the elements.

The mass is a linear combination as follows:

$$\|N_j \cdot M_j^t\| = \left\| \left(n_1, \cdots, n_j\right) \cdot \left(m_1 \ \cdots \ m_j\right)^t \right\| = \sum_{i=1}^{j} n_i \times m_i = m \tag{2.1}$$

with $N_j$, the Diophantine set of stoichiometry; $M_j$, the masses of the elements.

https://doi.org/10.1017/S1743921319008032 Published online by Cambridge University Press

*Change of basis.* The goal of this work is to introduce the chemical rules in the way the candidate solutions to a given Diophantine equation are generated. The nitrogen rule states that molecular ions containing exclusively hydrogen, carbon, nitrogen and oxygen have an *odd nominal mass* when an *even number* of nitrogen atoms are present and an *even nominal mass* when an *odd number* of nitrogen atoms are present. For radical ions, the nitrogen rule becomes reversed. Since the generation algorithm is an integer exhaustive search, we need first to demonstrate that the use of alternative basis satisfies the wanted constraints. Any Diophantine set in $\mathbb{N}^j$ can be written as a vector in an alternative stoichiometric space. Indeed, any set of $j$ linearly independent vectors in $\mathbb{Z}^j$ form a basis. The change of basis is built as follow from the canonical basis C,H,N,O to the chosen alternative basis C,$CH_2$,NH,O:

$$A = \begin{array}{c} \\ C \\ H \\ N \\ O \end{array} \begin{array}{c} \begin{array}{cccc} C & CH_2 & NH & O \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}, A^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.2}$$

$A$ is the decomposition of the new basis vectors on the canonical stoichiometric basis and $A^{-1}$ the inverse of $A$. Since the alternative basis is a set of $j$ linearly independent vectors, $A$ is always square and invertible. Let us write a stoichiometric decomposition in the C,$CH_2$,NH,O basis as $N'_j = \left(n'_C, n'_{CH_2}, n'_{NH}, n'_O\right)$. It comes:

$$N'^t_j = A^{-1} \cdot N^t_j = \begin{pmatrix} 1 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} n_C \\ n_H \\ n_N \\ n_O \end{pmatrix} = \begin{pmatrix} [l] n'_C \\ n'_{CH_2} \\ n'_{NH} \\ n'_O \end{pmatrix} \tag{2.3}$$

$$m = \|N'_j \cdot M'^t_j\| = \left(n'_C, n'_{CH_2}, n'_{NH}, n'_O\right) \cdot \left([l]12\ 14.0156\ 15.0109\ 15.9949\right)^t \tag{2.4}$$

*Natural number coefficients and parity chemical rules.* Few rules are commonly accepted to filter most probable molecular identification out of mass measurements (Kind & Fiehn (2007)). Our goal here is to avoid a heuristic approach and go through an exhaustive search that incorporates restrictions. The idea is to take advantage of the fact that most of the rules can be set when a $A$ matrix is chosen. The image of any Diophantine $N'$ set through the $A$ linear application is trivially a Diophantine set $N$. In other words, any combination of natural number of $\left(n'_C, n'_{CH_2}, n'_{NH}, n'_O\right)$ is a stoichiometric formula with positive integer coefficients. The reciprocal is not true; the $A^{-1}$ matrix in Eq. 2.2 has non integer coefficients. Hereafter, we demonstrate that parity properties are stable through natural number combination of the chosen groups.

The equivalence between having natural number coefficients and respecting the nitrogen rule relies on the reciprocal. Let us demonstrate that the decomposition on the $\left(n'_C, n'_{CH_2}, n'_{NH}, n'_O\right)$ basis of any stoichiometric Diophantine set $N_j$ respecting the nitrogen rule will have integer coefficients:

$$\begin{pmatrix} 1 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} n_C \\ n_H \\ n_N \\ n_O \end{pmatrix} = \begin{pmatrix} n_C - \frac{n_H}{2} + \frac{n_N}{2} \\ \frac{n_H}{2} - \frac{n_N}{2} \\ n_N \\ n_O \end{pmatrix} \tag{2.5}$$

https://doi.org/10.1017/S1743921319008032 Published online by Cambridge University Press

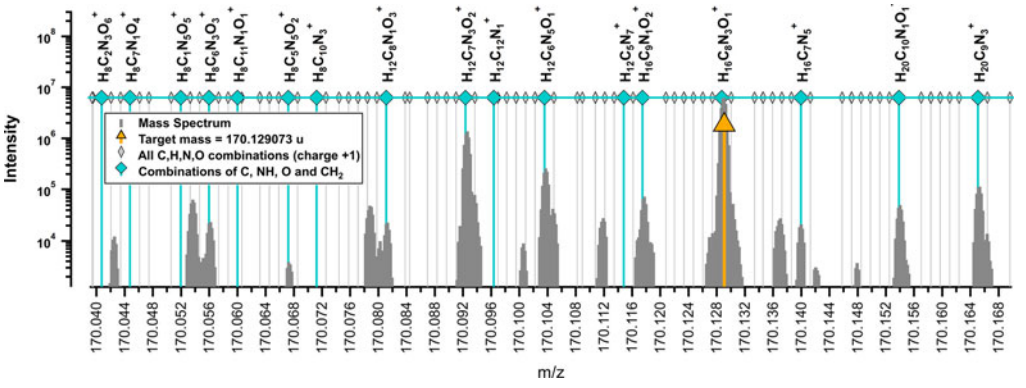


**Figure 2.** Comparison of an exhaustive search with C, H, N, O and C, $CH_2$, NH, O combination in a given mass range around 170.13u. The much higher density without change of basis is of no help close to the target mass since none of additional combinations has a better match than the one produced with C, $CH_2$, NH, O combinations.

if $n_N$ and $n_H$ are odd:

$$\begin{cases} n_N \equiv 1 \pmod 2 \\ n_H \equiv 1 \pmod 2 \end{cases} \Rightarrow \begin{cases} \frac{n_H}{2} \equiv \frac{1}{2} \pmod 1 \\ \frac{n_N}{2} \equiv \frac{1}{2} \pmod 1 \end{cases} \Rightarrow \left(\frac{n_H}{2} - \frac{n_N}{2}\right) \equiv 0 \pmod 1 \quad (2.6a)$$

$$n_C \equiv 0 \pmod 1 \Rightarrow \left(n_C - \frac{n_H}{2} + \frac{n_N}{2}\right) \equiv 0 \pmod 1 \quad (2.6b)$$

then if $n_N$ and $n_H$ are even:

$$\begin{cases} n_N \equiv 0 \pmod 2 \\ n_H \equiv 0 \pmod 2 \end{cases} \Rightarrow \begin{cases} \frac{n_H}{2} \equiv 0 \pmod 1 \\ \frac{n_N}{2} \equiv 0 \pmod 1 \end{cases} \Rightarrow \left(\frac{n_H}{2} - \frac{n_N}{2}\right) \equiv 0 \pmod 1 \quad (2.7a)$$

$$n_C \equiv 0 \pmod 1 \Rightarrow \left(n_C - \frac{n_H}{2} + \frac{n_N}{2}\right) \equiv 0 \pmod 1 \quad (2.7b)$$

Thus, respecting the nitrogen rule and having $N'_j \in \mathbb{Z}^j$ is equivalent.

*Coefficients positivity and stoichiometric ratio.* A remarkable property of this change of basis is that introducing constraints in the elemental ratios is equivalent to put constraints on the coefficients. This can be done by chosing only natural numbers for $N'_j$ coefficients. The chosen $A$ matrix has peculiar properties if $N'_j \in \mathbb{R}^{+j}$. Indeed:

$$\begin{cases} n_C - \frac{n_H}{2} + \frac{n_N}{2} \geqslant 0 \\ \frac{n_H}{2} - \frac{n_N}{2} \geqslant 0 \end{cases} \Rightarrow \begin{cases} n_H \leqslant 2n_C + n_N \Leftrightarrow \mathrm{DBE} \geqslant 1 \\ n_N \leqslant n_H \end{cases} \quad (2.8)$$

With this $A$ matrix, the positivity of the coefficients of $N'_j$ implies that the DBE (double bond equivalent, given by$1 + n_C - \frac{n_H}{2} + \frac{n_N}{2}$) cannot be lower than 1 and that the number of nitrogen cannot exceed the number of hydrogen atoms. An example is given in Figure 2.

## 3. Exhaustive search for a stoichiometric formula

Any natural numbers linear combination of $\left(n'_C, n'_{CH_2}, n'_{NH}, n'_O\right)$ will have DBE $\geqslant 1$, $n_N \leqslant n_H$ and will respect the nitrogen rule. We can now build an algorithm to search for a natural Diophantine set associated to the closest mass relative to a measurement peak. If the error model is $\hat{m} = m + \hat{e}$, with $\hat{e}$ being the measured error, there is no easy way to directly estimate the mass bias $m - \hat{m}$ without knowing the error *a priori*. Choice is made not to use a heuristic algorithm but an exhaustive search, to have a complete knowledge of the mass biases, the distances between computed masses ($m$) and measured mass ($\hat{m}$). The following sections describe how the set of computed masses are generated.

https://doi.org/10.1017/S1743921319008032 Published online by Cambridge University Press

**Table 2.** Coefficients array for the $(j-1)^{\text{th}}$ heaviest masses.

| | $n'_{CH_2}$ | $n'_{\chi}$ | $n'_{NH}$ | $n'_{O}$ |
|---|---|---|---|---|
| $\prod_{k=1}^{j-1}(max_k+1)$ rows { $(max_{j-2}+1)(max_{j-1}+1)$ rows { $(max_{j-1}+1)$ rows { | 0 | ⋯ | 0 | 0 ⋮ $max_{j-1}$ |
| | | | ⋮ | 0 ⋮ $max_j$ |
| | | | $max_{j-2}$ | 0 ⋮ $max_j$ |
| | 1 ⋮ $max_1$ | | 0 ⋮ $max_{j-2}$ | 0 ⋮ $max_{j-1}$ |

*The Diophantine equation itself*. There is no algorithm that allows to prove there is a solution to a Diophantine equation in general (Robinson (1972)). There are only few theorems for peculiar cases. The Bezout's identity sets one condition for a linear Diophantine equation to have integer solutions:

$$\forall(m_1,\cdots,m_j)\in\mathbb{Z}^j,\exists(n_1,\cdots,n_j)\in\mathbb{Z}^j,n_1\cdot m_1+\cdots+n_j\cdot m_j=\gcd(m_1,\cdots,m_j) \quad (3.1)$$

Where $\gcd(m_1,\cdots,m_j)$ is the greatest common divisor or $m_1,\cdots,m_j$. In the case of non-negative solutions, a solution exists for each number greater than the Frobenius number of the $(m_1,\cdots,m_j)$ set. For instance, 4 being the Frobenius number of $(3,5,7)$, 4 is the largest rugby score that cannot be obtained. There is no closed-form for $j\geqslant 3$. Frobenius number is defined if and only if the $m_j$ are mutually prime. For our basis groups, the rounded masses: $(12,14,15,16)$ are coprime integers and their Frobenius number is 49, which means every integer mass greater than 49 can be associated with a Diophantine set (Einstein *et al.* (2007)). Finding the Diophantine sets requires two steps: one to enumerate sets and one to check if they are solution to the equation.

*The $N'_{1\to j-1}$ enumerator*. The goal is to enumerate all the combinations of all masses but the lightest and to fit in the remainder with the lightest mass. It is an Euclidian-like algorithm. Let us call the $j^{th}$ mass the lightest one so :

$$m'_j<m'_1<\cdots<m'_{j-1} \quad (3.2)$$

Let us build the array shown in Table 2. To do so in a vectorised manner we need to calculate the range for the coefficient to span $MAX=(max_1,\cdots,max_{j-1})$ and the stretches $STR=(str_1,\cdots,str_{j-1})$ as follow:

$$max_k=\left\lfloor\frac{\hat{m}}{m'_k}\right\rfloor \quad \text{and} \quad str_k=\begin{cases}\prod_{i=k+1}^{j-1}(max_i+1) & \text{if} \quad k<j-1\\ 1 \quad \text{else}\end{cases} \quad (3.3)$$

With $\lfloor x\rfloor$ being the floor function of $x$. Then, with $\kappa$ being the 1-based column index and $\rho$ being the 0-based row index, the value in a cell of the array in Table 2 is:

$$T(\kappa,\rho,STR,MAX)=\left\lfloor\frac{\rho}{str_\kappa}\right\rfloor \bmod (max_\kappa+1) \quad (3.4)$$

https://doi.org/10.1017/S1743921319008032 Published online by Cambridge University Press

In our example of the $A$ matrix, one can compute the row where a $({n'}_{CH_2} \leqslant max_{CH_2}, {n'}_{NH} \leqslant max_{NH}, {n'}_O \leqslant max_O)$ set is:

$$\rho({N'}_{j-1}, STR) = \sum_{k=1}^{j-1} ({n'}_k \times str_k) \tag{3.5}$$

Let us remark a quick method to build a bijection between $\mathbb{N}^{j-1}$ and $\mathbb{N}$.

*Jumping the array: avoiding pointless calculation.* For each row, the $j-1$ heaviest mass coefficients are set. The lightest is still to be computed to approximate the $\hat{m}$ target:

$${n'}_j(\rho) = \left\lfloor \frac{\hat{m} - \sum_{k=1}^{j-1}({n'}_k \times {m'}_k)}{{m'}_j} \right\rceil = \left\lfloor \frac{\bar{m}}{{m'}_j} \right\rceil \tag{3.6}$$

Where $\lfloor x \rceil$ is the rounding function of $x$, returning the closest integer. If $\bar{m} \leqslant \frac{{m'}_j}{2}$, it means the set coefficients already exceed the target mass to approximate. The array has the property that as the row index increments and one coefficient goes back to zero, $\sum_{k=1}^{j-1}({n'}_k \times {m'}_k)$ is guaranteed to decrease. Indeed, each time a coefficient goes back to zero, the coefficient in the column before increments. Since the masses of the basis groups are sorted, the sum decreases necessarily. The best way to decrease $\sum_{k=1}^{j-1}({n'}_k \times {m'}_k)$ is to skip all the rows until the value decreases, that is to say, jump to the row where the non-zero coefficient on the rightmost column returns to zero. The identification of the most right non-zero coefficient can be done by testing nullity from right to left in a given row. Let be $k$, the column index for the coefficient we want to set to zero, the index $\rho_{togo}$ is computed by ceiling the number of blocks with equal coefficient in the $k-1$ column:

$$\rho_{togo} = str_{k-1} \times \left\lceil \frac{\rho_{massexceeded}}{str_{k-1}} \right\rceil \tag{3.7}$$

Each time a ${N'}_{j-1}$ is set, that means for each row, the coefficients that are going to change in the next row are known as well as $\bar{m}$, the rest of the mass to fill. That is to say each time a coefficient changes, one can write a new Diophantine equation and check for the existence of a solution. Let $k$ be the column index of the last coefficient to increment on the next row, then:

$$\lfloor \bar{m} \rceil = ({n'}_k \lfloor {m'}_k \rceil) + \cdots + ({n'}_j \lfloor {m'}_j \rceil) \Leftrightarrow \begin{cases} \lfloor \bar{m} \rceil \equiv 0 \pmod{\gcd(\lfloor {m'}_k \rceil, \cdots, \lfloor {m'}_j \rceil)} \\ ({n'}_k, \cdots, {n'}_j \in \mathbb{Z}^{j-k}) \end{cases} \tag{3.8}$$

This means that if the rest to fill is not a multiple of the greatest common divisor of the $(k, \cdots, j)^{th}$ rounded masses, the equation has no integers solution and therefore no natural numbers solution. If true, there is no point at iterating on the rows and it is better jump to the next $\lfloor \bar{m} \rceil$ that is found the exact same way we jumped to the next lighter one with the Eq. 3.7.

*The algorithm.* In order to take into account a charge bearer that may not respect the nitrogen rule and that we do not want to decompose, its mass has to be subtracted before computation. The constraints on the elemental ratios can be overcome by considering additional groups into the generator. For instance, if $N_2$ is added to the group set, the $n_N \leqslant n_H$ constraint is no more. This also can be done by manual setting of the ranges, at users risks.

https://doi.org/10.1017/S1743921319008032 Published online by Cambridge University Press

**Algorithm 1** Diophantine Solver

**Data:** $m$: mass to decompose; $M_j$: list of $j$ masses sorted like in eq.3.2
**Result:** $S_r$: list of approximations of $m$; $C_{r,j}$: coefficients array associated
**begin**
  Remove the assumed charge bearers mass (protons and electrons)
  Initialize empty $C$ and $S$
  Set the ranges and stretches values for each coefficient
  Compute the maximum number of rows
  Initialize the row counter
  **while** *row counter* $\leqslant$ *maximum number of rows* **do**
    **for** *all masses but the lightest* **do**
      Compute the coefficients (eq. 3.4)
    Find the non-zero coefficient associated with the heaviest mass
    **if** *There is no solution (eq. 3.8) OR Target mass is exceeded* **then**
      Jump to the next candidate row (eq. 3.7)
    **else**
      Complete with the lightest mass (eq. 3.6)
      Add the results to $C$ ans $S$
      Increment the row counter
  Adjust with the constant parameters removed at the beginning
  Sort the two output lists by absolute bias

## Acknowledgment

This work is supported by the French Space Agency (CNES) under their Exobiology and Solar System programs.

## References

Danger, G., Orthous-Daunay, F. R., de Marcellus, P., Modica, P., Vuitton, V., Duvernay, F., Flandinet, L., Le Sergeant d'Hendecourt, L., Thissen, R., & Chiavassa, T. 2013, *Geochim. Cosmochim. Acta*, 118, 184

Einstein, D., Lichtblau, D., Strzebonski, A., & Wagon, S. 2007, *E. J. Comb. Num. Th.*, 7, 63

Kind, T. & Fiehn, O. 2007, *BMC bioinformatics*, 8, 105

Kozhinov, A. N., Zhurov, K. O., & Tsybin, Y. O 2013, *Anal. Chem.*, 85, 6437

Makarov, A. 2000, *Anal. Chem.*, 72, 1156

Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., & Horning, S. 2006, *Anal. Chem.*, 78, 2113

Meija, J. 2006, *Anal. Bioanal. Chem.*, 385, 486

Pernot, P., Carrasco, N., Thissen, R., & Schmitz-Afonso, I. 2010, *Anal. Chem.*, 82, 1371

Robinson, J. 1972, *J. Symbolic Logic*, 37, 605

Yamashita, M. & Fenn, J. B. 1984, *J. Phys. Chem.*, 88, 4451

https://doi.org/10.1017/S1743921319008032 Published online by Cambridge University Press