Obtaining the Molecular Formula from its Percent Composition: an Algebraic Formulation

James B. Callis
Professor of Chemistry
University of Washington

Let us begin with a compound whose chemical formula can be represented as $A_aB_bC_c$, where each atom making up the compound is identified by its chemical symbol, e.g. A, and the number of atoms of type A in the compound is identified with a subscript, e.g. a, which represents a whole number. Thus, the compound under discussion has a atoms of type A, combined with b atoms of type B and c atoms of type C. For the set of numbers {a b c} we consider two possibilities: (i) the set has no common multiplier other than one, or (ii) the set has a whole number common multiplier, say δ, which allows us to represent the set as {a b c} as {δa' δb' δc'} = δ{a' b' c'}. Now if we define the symbol $A_aB_bC_c$ as the molecular formula, the formula $A_{a'}B_{b'}C_{c'}$ represents a different formula, the empirical formula. In the case where δ = 1, the molecular formula and the empirical formula are identical. However in the case where δ is a whole number greater than 1 then a', b' and c' are all smaller than a, b and c by the factor δ. Generally, we denote the case where a', b' and c' is the set containing the smallest possible whole numbers we have the chemical empirical formula.

As an example consider the molecular formula for glucose, $C_6H_{12}O_6$. This compound is characterized by the set {6 12 6}, which can be written {6·1 6·2 6·1} = 6{1 2 1}, i.e. δ = 6 and a' = c' = 1 with b =2. Thus we can state the empirical formula for glucose as CH_2O. Associated with each molecular and empirical formula is a molecular mass (MM) and empirical mass (EM). The former can be calculated as follows:

$$MM_{A_aB_bA_C} = a \cdot M_A + b \cdot M_B + c \cdot M_C \tag{1}$$

Where MM_{AaBbCc} is the molar mass (in grams) of the compound $A_aB_bC_c$ and M_A is the atomic mass in grams of the element A, and similarly for M_B and M_C. The calculation for the empirical mass is similar:

$$EM_{A_{a'}B_{b'}A_{c'}} = a' \cdot M_A + b' \cdot M_B + c' \cdot M_C \tag{2}$$

At this point we leave it to the student to show that

$$MM_{A_aB_bC_c} / EM_{A_{a'}B_{b'}C_{c'}} = \delta \tag{3}$$

With the above definitions in mind we are now ready to pose the problem stated in the title: *Find* the molecular formula of a compound *given* its fractional composition by mass together with the values of the atomic masses, *using* Dalton's atomic theory. To make the solution more concrete, we will assume the molecular formula above ($A_aB_bC_c$). Further, we will define the amount of material in a sample of the compound, $\text{mass}^{\text{sam}}_{A_aB_bC_c}$ as:

$$\text{mass}^{\text{sam}}_{A_aB_bC_c} = n^{\text{sam}}_{A_aB_bC_c} \cdot MM_{A_aB_bC_c} = n^{\text{sam}}_{A_aB_bC_c} \left(a \cdot M_A + b \cdot M_B + c \cdot M_C \right) \tag{4}$$

Where $n^{\text{sam}}_{A_aB_bC_c}$ is the number of molecules in the sample. The results of equation 4 can be expanded to include the possibility of an empirical formula that is different than the molecular formula.

$$\text{mass}^{\text{sam}}_{A_aB_bC_c} = n^{\text{sam}}_{A_aB_bC_c} \cdot \delta \cdot EM_{A_{a'}B_{b'}C_{c'}} = n^{\text{sam}}_{A_aB_bC_c} \cdot \delta \left(a' \cdot M_A + b' \cdot M_B + c' \cdot M_C \right) \tag{5}$$

We see that if $\delta = 1, \quad EM_{A_{a'}B_{b'}C_{c'}} = MM_{A_aB_bC_c}$ and the results are unambiguous. However, if $\delta > 1$,

$$\delta \cdot EM_{A_{a'}B_{b'}C_{c'}} = MM_{A_aB_bC_c} \tag{6}$$

Let us now evaluate the fractional composition of element A, f_A in the sample of $A_aB_bC_c$.

$$f_A = \frac{n^{\text{sam}}_{A_aB_bC_c} \cdot a \cdot M_A}{n^{\text{sam}}_{A_aB_bC_c} \cdot MM_{A_aB_bC_c}} = \frac{a \cdot M_A}{MM_{A_aB_bC_c}} \tag{7}$$

In passing we note that the formulation of the problem as a ratio in equation 7 has eliminated one of the variables, $n^{\text{sam}}_{A_aB_bC_c}$. We no longer need be concerned with this variable. As was done in equation 5, we expand equation 7 to include the possibility of an empirical formula whose coefficients are smaller than those in the molecular formula.

$$f_A = \frac{n^{\text{sam}}_{A_aB_bC_c} \cdot \delta \cdot a' \cdot M_A}{n^{\text{sam}}_{A_aB_bC_c} \cdot \delta \cdot EM_{A_{a'}B_{b'}C_{c'}}} = \frac{a' \cdot M_A}{EM_{A_{a'}B_{b'}C_{c'}}} \tag{8}$$

We see that owing to the division operation we have removed δ from the equation. This implies that we have lost the ability to determine the molecular formula and can only recover the empirical formula. Similarly, we may derive expressions for f_B and f_C:

$$f_B = \frac{b' \cdot M_B}{EM_{A_{a'}B_{b'}C_{c'}}}; \quad f_C = \frac{c' \cdot M_C}{EM_{A_{a'}B_{b'}C_{c'}}} \tag{9}$$

Equations 8 and 9 form a system of three equations in ten variables: f_A, f_B, f_C, $EM_{A_{a'}B_{b'}C_{c'}}$, M_A, M_B, M_C, a’, b’, and c’. Of these variables, six are known quantities (f_A, f_B, f_C, M_A,

M_B, M_C) and four are unknown (a', b', c', $EM_{A_{a'}B_{b'}C_{c'}}$). As currently stated, this problem is underdetermined. Moreover, we seem to have eliminated the ability to determine the molecular formula because neither it nor δ is present in the set of equations.

Nevertheless, we continue. Our next manipulation is to eliminate the known the mass of each element that is present in one mole of sample. This is accomplished by dividing each fraction by its corresponding atomic mass. Equations 8 and 9 become:

$$\frac{f_A}{M_A} = \frac{a'}{EM_{A_{a'}B_{b'}C_{c'}}}; \quad \frac{f_B}{M_B} = \frac{b'}{EM_{A_{a'}B_{b'}C_{c'}}}; \quad \frac{f_C}{M_C} = \frac{c'}{EM_{A_{a'}B_{b'}C_{c'}}}; \qquad (10)$$

At this point, we have a system of three equations in 7 variables: f_A, f_B, f_C, $EM_{A_{a'}B_{b'}C_{c'}}$, a', b', and c'. Of these variables, three are known quantities (f_A, f_B, f_C,) and four are unknown (a', b', c', $EM_{A_{a'}B_{b'}C_{c'}}$). To simplify the problem further, we eliminate the empirical mass by dividing each of the equations by the equation with the smallest value. Suppose this is the first equation in 10. Then we have:

$$\frac{\frac{f_A}{M_A}}{\frac{f_A}{M_A}} = \frac{\frac{a'}{EM_{A_{a'}B_{b'}C_{c'}}}}{\frac{a'}{EM_{A_{a'}B_{b'}C_{c'}}}} = 1 \qquad (11)$$

$$\frac{\frac{f_B}{M_B}}{\frac{f_A}{M_A}} = \frac{\frac{b'}{EM_{A_{a'}B_{b'}C_{c'}}}}{\frac{a'}{EM_{A_{a'}B_{b'}C_{c'}}}} = \frac{b'}{a'} \qquad (12)$$

$$\frac{\frac{f_C}{M_C}}{\frac{f_A}{M_A}} = \frac{\frac{a'}{EM_{A_{a'}B_{b'}C_{c'}}}}{\frac{a'}{EM_{A_{a'}B_{b'}C_{c'}}}} = \frac{c'}{a'} \qquad (13)$$

We see that taking the ratio of ratios has eliminated equation 11 (it is now trivial) and reduced the problem to one of a system of two equations (12 and 13) in three unknowns (a' b' c'). We have now arrived at the simplest possible expression of the problem as one that is ill-posed. Fortunately there is a further constraint that we can require for the solution that reduces the problem to one that is well-posed. This constraint may be stated as follows:

$$\{a'\ b'\ c'\}_{\text{min}}^{\text{whole nos.}} \qquad (14)$$

Thus the coefficients, {a’ b’ c’} must be such that they form the minimum set of whole numbers that satisfy equations 12 and 13. We can also restate equation 14 in terms of a Cartesian space with axes along the coefficient directions. In this case we are asking for the point given by coordinates {a’ b’ c’} that express that the minimum set of coefficients is the lowest possible distance from the origin in the Euclidean sense, i.e. that

$$\left\{(a')^2 + (b')^2 + (c')^2\right\}_{\text{min}}^{\text{whole nos.}} \qquad (15)$$

To provide a concrete illustration of the concepts in this work, we consider the case of adipic acid. The molecular formula for this compound is $C_6H_{10}0_4$, MM = 146.14 g/mol. The empirical formula is $C_3H_50_2$, EM = 73.07 g/mol and $\delta = 2$. With this data we can calculate the fractional elemental abundances: C = 0.4931, H = .06897 and O = 0.4379. We are now ready to pose the central problem as applied to adipic acid: *Find* the empirical formula of adipic acid *given* its fractional composition by mass together with the values of the atomic masses, *using* Dalton’s atomic theory.

Our first step is to eliminate the number of molecules in the original sample by forming the following ratios using the definitions found in equation 10:

$$\frac{f_C}{M_C} = \frac{.4931}{12.01} = .04106; \quad \frac{f_H}{M_H} = \frac{.06897}{1.007} = 0.06932; \quad \frac{f_O}{M_O} = \frac{.4379}{16.00} = .02737 \qquad (16)$$

Next we eliminate the empirical mass from the equations by dividing by the lowest value of f_X/M_X. In this case it is $f_O/M_O = .02737$. Using the definitions found in equations 11 – 13, we find the following ratios:

$$\frac{\frac{f_O}{M_O}}{\frac{f_O}{M_O}} = \frac{.02737}{.02737} = 1; \quad \frac{\frac{f_C}{M_C}}{\frac{f_O}{M_O}} = \frac{.04106}{.02737} = 1.5002; \quad \frac{\frac{f_H}{M_H}}{\frac{f_O}{M_O}} = \frac{.06932}{.02737} = 2.532 \qquad (17)$$

The above set of numbers is identical within experimental and round off error to the set of rational numbers {1 3/2 5/2}. These can then be converted to the smallest set of whole numbers by multiplication by 2 (least common multiplier). This final set of whole numbers is identical to the empirical formula of adipic acid that we started with.

Our final task is to show how to go from the empirical formula to the molecular formula. Equation 3 holds the key. If we have the set {a’ b’ c’} and the empirical molar mass, we need only one more parameter. According to equation 3 it may be either δ or $MM_{A_aB_bC_c}$. Experimentally, it is usually easier to obtain $MM_{A_aB_bC_c}$, which then yields δ.