A Digital Tool for Liquid-Liquid Extraction Process Design

Authors:

George Karageorgis*[a], Simone Tomasi[a], Elliot Farrar[a], Maxime Tarrago[a] , Tabassum Malik[a]

Supplementary Information

Affiliations:

a: Chemical Development, Pharmaceutical Technology & Development, Operations, AstraZeneca, Macclesfield, U.K.

Table of Contents

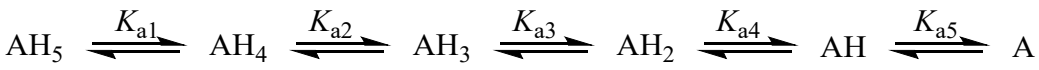
General

The code and data associated with this manuscript can be accessed through a publicly available repository on GitHub:

https://github.com/AstraZeneca/LLE_Digital_Tool

1. Equations describing extract processes of ionisable compounds

We use a generalised mass balance equation describing the distribution of the molar fraction f_i of each ionic species of a compound across the range of the pH scale (0-14, Eq. 1) as described previously.[1] The determination of f_N in the case of compounds with multiple ionic species is crucial and not trivial. However, we can consider the example of a compound with 5 pK_a values as follows:



Scheme S1: Equilibria of the forms of a compound with 5 ionisable positions.

A compound with n ionisable positions will have n+1 forms. In this example the charge of each form decreases moving left to right (from AH_5 to A), allowing us to identify the neutral form of the compound. The molar fraction of each of the above forms can be calculated using the following form equations:

$$f_{AH_5} = \frac{1}{1 + \frac{K_{a1}}{[H^+]} + \frac{K_{a1}K_{a2}}{[H^+]^2} + \frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}$$

Form Equation S1

$$f_{AH_4} = \frac{\frac{K_{a1}}{[H^+]}}{1 + \frac{K_{a1}}{[H^+]} + \frac{K_{a1}K_{a2}}{[H^+]^2} + \frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}$$

Form Equation S2

[1] Ashworth, I. W.; Meadows, R. E. A General Liquid–Liquid Partitioning Equation and Its Consequences: Learning from the pH Dependent Extraction of a Pharmaceutical Intermediate. *The Journal of Organic Chemistry* 2018, 83 (7), 4270-4274. DOI: 10.1021/acs.joc.8b00309.

$$f_{AH_3} = \frac{\frac{K_{a1}K_{a2}}{[H^+]^2}}{1 + \frac{K_{a1}}{[H^+]} + \frac{K_{a1}K_{a2}}{[H^+]^2} + \frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}$$

Form Equation S3

$$f_{AH_2} = \frac{\frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3}}{1 + \frac{K_{a1}}{[H^+]} + \frac{K_{a1}K_{a2}}{[H^+]^2} + \frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}$$

Form Equation S4

$$f_{AH} = \frac{\frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4}}{1 + \frac{K_{a1}}{[H^+]} + \frac{K_{a1}K_{a2}}{[H^+]^2} + \frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}$$

Form Equation S5

$$f_{A} = \frac{\frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}{1 + \frac{K_{a1}}{[H^+]} + \frac{K_{a1}K_{a2}}{[H^+]^2} + \frac{K_{a1}K_{a2}K_{a3}}{[H^+]^3} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}}{[H^+]^4} + \frac{K_{a1}K_{a2}K_{a3}K_{a4}K_{a5}}{[H^+]^5}}$$

Form Equation S6

The molar fraction of the compound present in neutral form f_N can be calculated by the equation which corresponds to the neutral form of the compound. For example, if the neutral form is AH, f_N can be calculated using Form Equation S5.

The above form equations can be generalised in Equation 1

$$f_i = \frac{\prod_{i=0}^{N} K_{a,j} \Big/ [H^+]^j}{\sum_{i=0}^{N} \left(\prod_{i=0}^{N} K_{a,j} \Big/ [H^+]^j \right)}$$

Equation 1

where: $$K_{a,j} = \begin{cases} j = 0 \rightarrow 1 \\ j > 0 \rightarrow K_{a,j} \end{cases}$$

In Equation 1, N is the number of dissociation constants of the solute. $K_{a,j}$ is the dissociation constants of the compound in decreasing order (increasing pK_a order). Using Equation 1, we can also calculate the molar fraction extracted into either the aqueous f_{aq} or the organic phase f_{org} for each compound as follows:

$$f_{aq} = \frac{1}{1 + K_P V_R f_N} \qquad \text{Equation 2}$$

$$f_{org} = 1 - f_{aq} \qquad \text{Equation 3}$$

In Equation 2, f_{aq} is the mole fraction of the solute extracted into the aqueous phase, K_P is the partition coefficient of the neutral species, V_R is the volume ratio, defined as $V_R = V_{org}/V_{aq}$, and f_N is the molar fraction of the compound present in the neutral form.

Finally, we can define extraction efficiency. In the case of just a single compound, the extraction efficiency represents the fraction of the compound. This can be shown as follows:

To begin, the fraction of the compound in the extract and raffinate phases are complementary:

$$f_{extract}^{comp} + f_{raffinate}^{comp} = 1 \qquad \text{Equation S4}$$

The extraction efficiency represents the fraction of the compound to isolate that moves to the extract compared to the fraction of the compound that moves to the raffinate at the end of the extraction.[2] Mathematically, it can be expressed as:

$$Eff_{ext}^{comp} = \left\{ 1 + \left[\frac{f_{ext}^{comp} - f_{raf}^{comp}}{f_{ext}^{comp} + f_{raf}^{comp}} \right] \right\} \Big/ 2$$

From S4 we can replace $f_{raffinate}^{comp}$:

$$Eff_{ext}^{comp} = \left\{ 1 + \left[\frac{f_{ext}^{comp} - \left(1 - f_{ext}^{comp}\right)}{f_{ext}^{comp} + \left(1 - f_{ext}^{comp}\right)} \right] \right\} \Big/ 2$$

[2] Müller, E., Berger, R., Blass, E., Sluyts, D. and Pfennig, A. Liquid–Liquid Extraction. In *Ullmann's Encyclopedia of Industrial Chemistry*, 2008, DOI: 10.1002/14356007.b03_06.pub2

and by simplification:

$$Eff_{ext}^{comp} = f_{ext}^{comp}$$

When an impurity is present, its distribution across the two phases can be similarly represented:

$$f_{ext}^{imp} + f_{raf}^{imp} = 1 \quad \text{Equation S5}$$

In this case the extraction efficiency is the product of the efficiency of the recovery of the compound and the efficiency of the rejection of the impurity:

$$Eff = Eff_{ext}^{comp} \times Eff_{ext}^{imp}$$

where:

$$Eff_{ext}^{imp} = 1 - f_{ext}^{imp} = f_{raf}^{imp}$$

As such:

$$Eff = f_{ext}^{comp} \times f_{raf}^{imp}$$

In the case of multiple impurities present, the extraction efficiency can be calculated as the product of the molar fraction of compound in the extract phase multiplied by the mean of the molar fractions rejected of all the impurities.[2] For example, if we are monitoring the organic phase, the extraction efficiency can be mathematically expressed as:

$$Eff_{org}^{comp} = f_{org}^{comp} \times \frac{\sum_{i=1}^{N} f_{aq}^{imp_i}}{N} \quad \text{Equation 4}$$

In Equation 4, N is the number of impurities, and the molar fractions of the isolatable compound f^{comp}_{org} or the impurities f^{imp}_{aq} in the respective phases can be calculated using Equations 2 and 3.

2. Table of LogP values

The LogP values used for this study were predicted with the COSMOtherm program, version 2024, (35) using *cosmo* files of BP-TZVPD-FINE quality. DFT calculations to generate the cosmo files were performed using the Turbomole package, Version 7.8[3] with the Becke−Perdew (BP) exchange−correlation functional[4,5] and Ahlrichs-type basis sets. Geometries were optimized using the def2-TZVP basis set (TZVP)[6] and energies and charges refined with single point calculations with the def2-TZVPD basis set and a fine marching grid (TZVPD-FINE). *Cosmo* file generation for solutes and a few solvents used an internally developed version of the ReSCoSS workflow (40), which includes conformational sampling and refinement, whereas *cosmo* files of most solvents were obtained from the database provided by the vendor with the COSMOtherm software. LogP predictions at infinite dilution and 25 °C were produced using the BP_TZVPD-FINE_C030_2016 parameterization, using the *eq_phases* keyword to set the composition of each aqueous/organic system to the LLE values, as predicted by COSMOtherm.

See file “1_2_3_pred_logP_data.csv”

[3] TURBOMOLE V7.3 2018, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.; (accessed 09/09/2024)
[4] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38* (6), 3098-3100. DOI: 10.1103/PhysRevA.38.3098
[5] Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B* **1986**, *33* (12), 8822-8824. DOI: 10.1103/PhysRevB.33.8822
[6] Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *PCCP* **2005**, *7* (18), 3297-3305, 10.1039/B508541A. DOI: 10.1039/B508541A

3. Table of pKa values of compounds.

See file “1_2_3_pKa_data.csv”

Column headers: The column names follow a naming convention of the prediction software used. A compound can have up to 4 “basic” (ACD_PKA_B[1-4]_PRED) and/or up to 4 “acidic” (“ACD_PKA_A[1-4]_PRED”) pKa values.

A basic pKa value denotes the loss of a proton to go from a positively charged state to a neutral state *e.g.* RNH_3^+ -> RNH_2 .

Conversely, an acidic pKa value denotes the acquisition of a proton to go from a negatively charged state to a neutral state *e.g.* RCO_2^- -> RCO_2H .

If a compound has one basic and one acidic pKa value, it means that the neutral form is the one in between the two pKa values.

If a compound has only two 2 basic pKa values, it means that the neutral form is the one after the second pKa value and that the first form is doubly positively charged.

If a compound has only 3 acidic pKa values, it means that the neutral form is the one before the first pKa value and the last form is triply, negatively charged.

4. Table of compounds’ SMILES.

See file “1_2_3_structures.csv”

5. Table of Solvent Physical Properties.

See file “solvents_full__wide_df.csv”

6. Example of python code use.

We have setup a public repository which contains the data and an example of use.

You can access it here: https://github.com/AstraZeneca/LLE_Digital_Tool

If you would like to run the jupyter notebook “example.ipynb” clone the repository, create a new virtual environment using Python 3.10 (or above) and install the required packages using the “LLE_Digital_Tool_requirements.txt”. If you would like to replicate the results included in the main text, you can follow the instructions provided in the “README.md” file.

If you would like to use the code with your own data, you will need to add the relevant data for your compounds to the appropriate csv files. Alternatively, you can adapt the code provided in the “LLEFunctions.py” file to conform with your data structures and objects. You could then edit the *system_panel_values* and *compounds_panel_values* objects to match your case.

7. Example of tool output.

See file “Results_PRODUCT 3_AMINE2_BROMDE1.csv”