ROYAL SOCIETY OF CHEMISTRY

Check for updates

A digital tool for liquid–liquid extraction process design†

George Karageorgis, [ID] * Simone Tomasi, [ID] Elliot H. E. Farrar, [ID] Maxime Tarrago and Tabassum Malik

Aqueous liquid–liquid extractions are crucial for purifying compounds and removing impurities in the pharmaceutical industry. However, the extensive solvent space involved in such operations highlights the need for an informed approach in solvent selection. We present a digital tool designed to leverage data-driven experimentation to enhance process efficiency and sustainability, aligning with industry trends towards digitalisation. It allows users to input various parameters, retrieve relevant data, and visualise extraction efficiencies, thereby improving process understanding and reducing process development lead times. By providing interactive visualisations and facilitating rapid hypothesis generation, the tool supports informed decision-making and streamlines workflows. The tool's application is demonstrated through representative complex scenarios involving the separation of multiple compounds present in a mixture at the end of a Buchwald coupling reaction. Overall, this digital tool offers a new practical and data-led approach to chemical process design, with the potential to promote experimental efficiency during development and to improve the environmental sustainability of commercial processes.

Introduction

Aqueous liquid–liquid extractions are a well-established workup and purification technique commonly used in many pharmaceutical processes. They are essential for removing key impurities, byproducts, undesired solvents or additives, while limiting product losses and maximising isolation efficiency. Optimal extraction conditions can afford more environmentally sustainable, leaner and cost-effective processes by reducing waste and process mass intensity (PMI), energy consumption, and operation time.[1,2] These improvements may also result in more cost efficient processes. As such, aqueous extractions play a crucial role in the pharmaceutical and fine chemical industry, significantly impacting both industrial processes and lead times.[3]

Across the industry, the drive towards digitalisation and standardisation in the development of chemical processes is more pertinent than ever, as it aligns with the broader goals of sustainability and efficiency.[4] Broadly speaking, this involves automatically collecting, analysing and displaying relevant data in a systematic fashion. The data is then leveraged to design targeted experimentations. This data-led approach can result in

substantial reductions in lead times for development and enhance the overall sustainability of chemical processes.[5,6] This digitally enabled way of working is facilitated by the development of tools that cater to non-experts. Such tools not only streamline workflows but also enable more informed decision-making through the integration of comprehensive data analysis. Here we introduce a digital tool which facilitates aqueous extraction design, exemplifying the benefits of such advancements.

The tool is designed to enable users to quickly investigate and perform virtual screenings of aqueous liquid–liquid extraction conditions. It enhances process understanding by providing interactive visualisations guiding users through targeted experimentation. Users are provided with a list of prioritised experiments, allowing them to focus on the most promising conditions. It represents a significant improvement compared with empirical ways of working, through offering a platform for rapid hypothesis generation, and standardised and efficient comparison of different extraction conditions. It allows for fast yet rational prioritisation of experimental outputs, even in time-sensitive, late-stage development scenarios.

Chemical Development, Pharmaceutical Technology & Development, Operations, AstraZeneca, Macclesfield, UK. E-mail: george.karageorgis@astrazeneca.com

† Electronic supplementary information (ESI) available: Equations describing extractive processes of ionisable compounds. Table of LogP values of compounds. Table of pK_a values of compounds. Table of compounds' SMILES. Table of Solvent Physical Properties. Example of Python code use. Example of tool output. See DOI: https://doi.org/10.1039/d5dd00104h

Results and discussion

Physicochemical basis

The mathematical expressions describing the physicochemical partitioning equilibrium of organic molecules are well established.[7,8] More recently, a general liquid–liquid partitioning

equation has been reported allowing the description of the distribution of organic materials with multiple ionic forms across two phases.[9] Briefly we use a generalised mass balance equation describing the distribution of the fraction f_i of each ionic species of a compound across the range of the pH scale (0–14, eqn (1), see ESI, Section 1† for further details).

$$f_i = \frac{\prod_{i=0}^{N} K_{a,j} / [H^+]^j}{\sum_{i=0}^{N} \left(\prod_{i=0}^{N} K_{a,j} / [H^+]^j \right)} \tag{1}$$

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

In eqn (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 eqn (1), we can also calculate the fraction extracted into either the aqueous f_{aq} or the organic phase f_{org} for each compound as follows:

$$f_{aq} = \frac{1}{1 + K_P V_R f_N} \tag{2}$$

$$f_{org} = 1 - f_{aq} \tag{3}$$

In eqn (2), 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 fraction of the compound present as the neutral form, in the aqueous phase.

Finally, we can define the extraction efficiency as the product of the fraction extracted of the isolated compound into the desired phase multiplied by the mean of the sum of the fractions rejected of all the impurities (see ESI, Section 1† for derivation and further details).[10] If we are monitoring the organic phase, the extraction efficiency can be mathematically expressed as:

$$\text{Eff}_{org}^{comp} = f_{org}^{comp} \times \frac{\sum_{i=1}^{N} f_{aq}^{imp_i}}{N} \tag{4}$$

In eqn (4), N is the number of impurities, and the fractions of the isolatable compound f_{org}^{comp} or the impurities f_{aq}^{imp} in the respective phases can be calculated using eqn (2) and (3).

Although the above equations represent a general solution to calculating the extraction efficiency of a compound in the presence of others there are still limitations to its implementation. Using the partition coefficient of a pure compound may result in deviations when applying the above equations to calculate the extraction efficiency of the same compound in the presence of other electrolytes or ion pairs, other organic compounds or at different temperatures.

Coding implementation

The above expressions lend themselves readily to digitalisation. We have done so using the Python programming language, allowing us to deploy the tool within AstraZeneca's intranet as an accessible web application using our internal scientific computing platform infrastructure. When designing the tool, we aimed to lower the accessibility barrier to our database for non-experts. To this end, the tool interface only requires the user to input basic information about the extractive process to be optimised, such as phases volumes, aqueous phase pH, which phase is the extractive, and which compounds are involved. Following essential validation, this information is used to construct standardised data queries sent to our database which retrieve the necessary data and store them in appropriate data objects. The data are processed and transformed to an appropriate format to allow faster calculations by the downstream functions. In short, the details of the extractive process and the details of the compounds involved in the separation are stored in Python dictionaries with several of their keys and value pairs updated as the necessary data is retrieved. We are sharing a fully functioning worked example of code which includes the data and procedural processing for the generation of the results discussed herein (see ESI,† Example Code). By sharing this code, we are enabling others to implement it and deploy it as it best suits them, on a platform of their choice.

Tool features

The tool works as illustrated in Fig. 1. The interface accepts various required user inputs such as organic and aqueous phases volumes, aqueous phase pH, organic solvent name, which phase the isolation is occurring in, and unique compound identifiers. If multiple compounds are entered, the user is required to specify which is the one they are interested in isolating or separating from the others. Once these data are entered, the tool queries our database for the required information: LogP values in a list of solvents and pK_a values for each compound. Once these physical properties data are retrieved, processed and validated, the tool produces a series of helpful visualisations that facilitate process understanding and allow further experimentation (Fig. 2). Such visualisations include the aqueous speciation of all compounds, pH-dependent fraction extracted plots for all compounds in the aqueous or the organic phase, as well as the extraction efficiency of the compound of interest. In each section, the user is provided with informational, confirmational, or warning messages facilitating their navigation and enhancing their interaction with the tool.

Application example

As a specific example of the use of this tool, we investigated two scenarios based on the Buchwald–Hartwig coupling of aryl bromide **1** with 4-methyl piperazine **2** to afford product **3** (Scheme 1), as previously reported.[11] This reaction represents a typical example where a liquid–liquid extraction step is required as part of the process and a case where a compound with multiple ionic forms such as **3**, can present a challenge for process modelling. The first scenario involves understanding the conditions required to efficiently separate an excess of the amine **2** from product **3**. The second scenario is concerned with the possible contingencies in the event where the reaction is
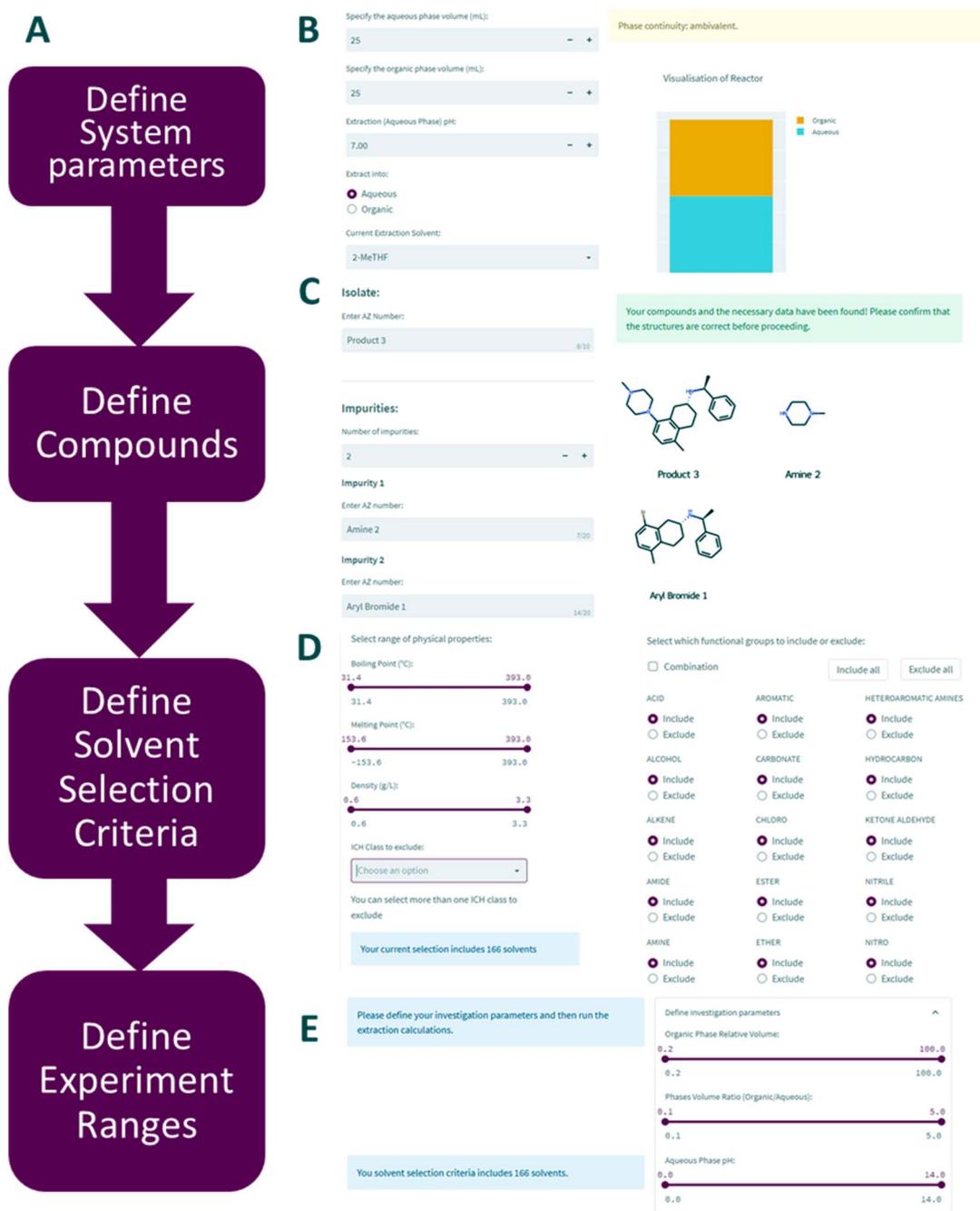
Fig. 1 Overview of the tool. Panel A: flowchart of actions. Panel B: input fields capturing information for the process. Panel C: input fields for the compounds involved. Visualisation allows structural verification. Panel D: the user can define a range of physical, regulatory and functional group criteria for solvent selection. Panel E: further definition of the operational ranges of the user's investigation. Informational (blue), confirmational (green), and warning (orange) messages are displayed in each section to enhance user experience.

incomplete, leading to a case where the crude reaction mixture contains a quantity of both starting materials **1** and **2**, as well as a significant amount of product **3.**

Scenario 1

In the published procedure, an excess of the amine **2** is used for the quantitative preparation of product **3** from starting material **1.** As such, upon reaction completion, a substantial amount of

amine **2** is left unreacted and requires separation from the desired product **3.** Using the unique identifiers, the user can immediately access speciation curves and fraction extracted graphs for each of these two compounds, as well as the extraction efficiency of the desired compound, product **3** (Fig. 2). The fraction extracted graphs can represent the fraction extracted either in the organic or the aqueous phase depending on which one the user is interested in. Here, both product **3** and amine **2**
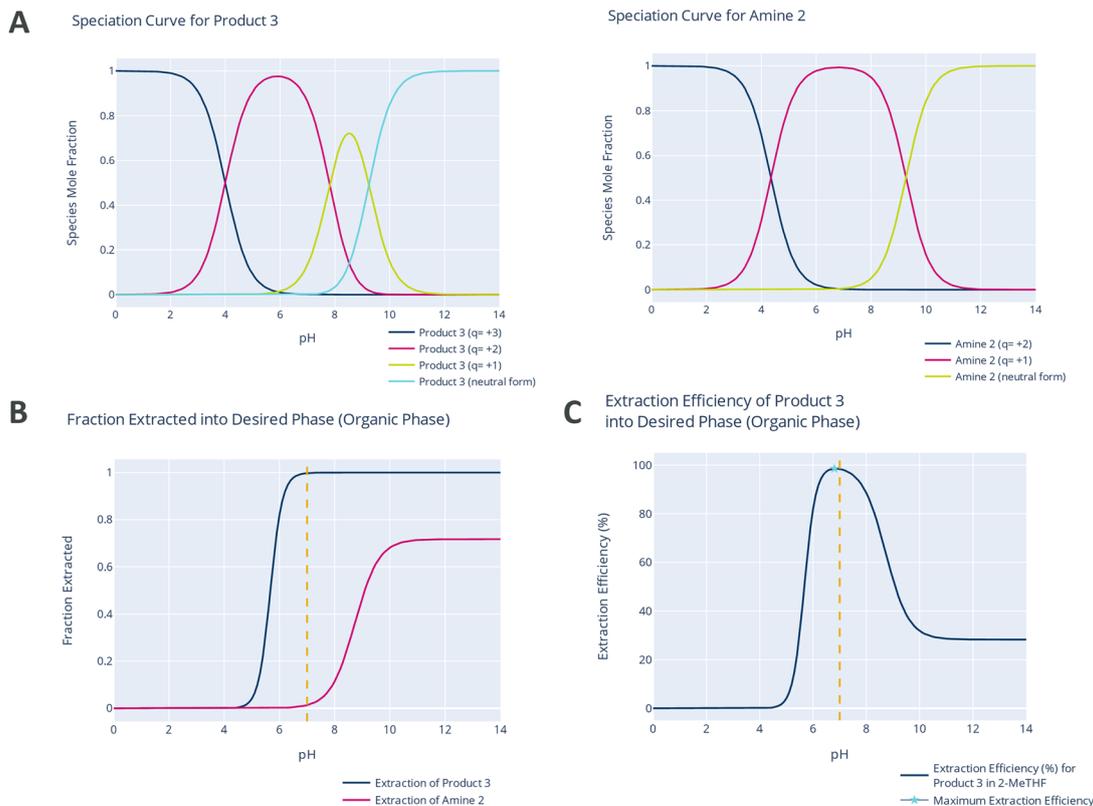
Fig. 2 Informative data visualisations for the process details provided. Panel A: speciation curves for product **3** (left) and amine **2** (right). Panel B: fraction extracted curves for both materials in the organic phase. Panel C: extraction efficiency of product **3** in the organic phase. The pH value for the maximum extraction efficiency is indicated with a star symbol. Volume ratio (organic/aqueous) = 1. Organic phase: 2-MeTHF.
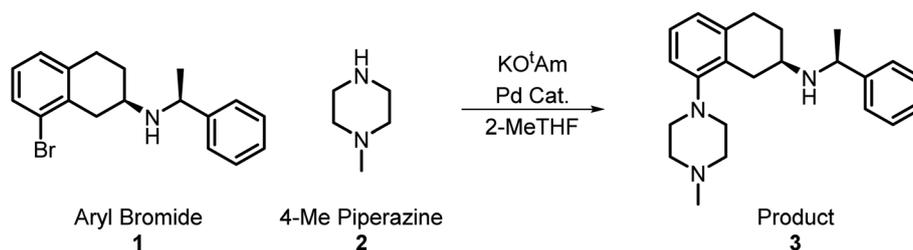

Scheme 1 Buchwald–Hartwig C–N coupling of starting materials **1** and **2** affording product **3**.

are protonated at very low pH, preventing them from extraction into the organic phase. At higher pH, both compounds' fraction extracted increase due to the decreasing ionised fraction of both species. Interestingly, the fraction extracted of **2** and **3** react differently to the pH for two reasons. First, the pK_a of **2** is lower than that of **3**, therefore the ionic fraction of the former is higher at a given pH. Second, even at high pH when both molecules are neutral, the fraction extracted of **2** is lower than the product **3**, owing to the lower LogP value of **2** in 2-MeTHF and water compared to product **3** (Scheme 2). Taken together, these observations highlight the existence of a "sweet spot" where optimising the extraction efficiency of **3** is feasible. Using the fraction extracted curves, the extraction efficiency of product **3** in the organic phase is automatically calculated by the tool

over the full pH range and displayed to the user (Fig. 2C). The pH corresponding to the optimal extraction efficiency is also highlighted. From these graphs, it is immediately evident that product **3** can be readily isolated from an equimolar amount of amine **2** by adding an equivolume amount of an aqueous phase with a suitable additive to adjust the pH to 7, in a high yield, directly from the crude reaction mixture. This conclusion reflects the experimental conditions included in the original report[11] where the necessary volume of water followed by a quantity of acetic acid are added to adjust the pH upon reaction completion.
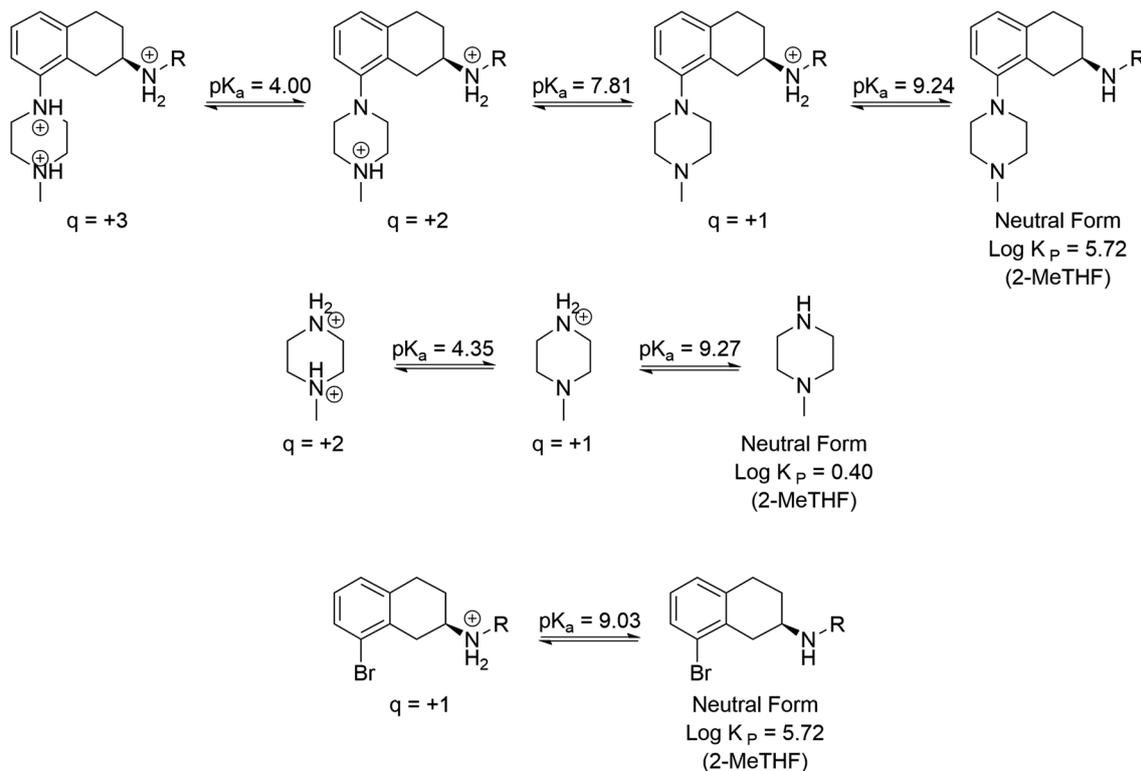
Scheme 2 Ionic and neutral forms for product **3** (top row), amine **2** (middle) and aryl bromide **1** (bottom). In each case the charge (q) for each ionic form is shown, and the logarithmic value for the partition coefficient (Log K_P) in 2-MeTHF for each compound's neutral form is given.

Scenario 2

In this scenario, we assume that the reaction has not progressed to completion, resulting in an end of reaction mixture containing unreacted aryl bromide **1**, excess amine **2**, and the formed product **3**. The tool can be used to retrieve the relevant data for all three compounds. As above, the speciation and fraction extracted curves can be used to gain insights into the opportunities available (Fig. 3). In this case, the fraction extracted curve of product **3** lies between the corresponding ones for the aryl bromide **1** and amine **2**. As such the isolation of product **3** in either the aqueous or the organic phase will be challenging. This observation is reflected in the extraction efficiency curves where the maximum extraction efficiency in either the aqueous or the organic phase does not exceed 50%. Notably the maximum extraction efficiency of product **3** in the aqueous phase is observed at a different pH value compared to that in the organic phase. This observation can be rationalised by considering the differences in the number of ionic forms and partition coefficient of the neutral form between the three compounds (Scheme 2).

In this scenario it is evident that product **3** cannot be efficiently separated from the other two compounds with one extraction. However, an experienced process chemist can resort to performing two sequential orthogonal extractions; one to separate the residual bromide **1** from the crude mixture, followed by one to separate amine **2** from product **3**. The organic solvent in the two extractions does not need to be the same;

screening for optimal solvents in both extractions could improve overall process efficiency and sustainability. The tool allows the user to define a set of physical criteria such as ranges of boiling points, melting points or densities, regulatory criteria such as ICH Classification,[12] or chemical functional groups, allowing the selection of solvents for subsequent calculations and comparisons. To this end, the user may also define the range of concentrations that are acceptable (expressed as process chemist-friendly relative volumes) and volume ratios. Solubility data is used to determine the minimum organic solvent volume required. All the retrieved and calculated data, as well as the investigation parameters and solvent selection criteria, can be exported in a FAIR data compatible format (comma separated files).[13,14] In this manner the data generated can be readily stored, retrieved, or re-used in other platforms, for example Microsoft Excel, allowing efficient knowledge retention and future reference.

In the event where multiple impurities are defined by the user, the tool is set to calculate the extraction efficiency of the desired compound against each of the impurities individually, accounting for this additional challenge. As shown in Fig. 4, an extraction with an equivolume quantity of an aqueous phase with a suitable additive to adjust the pH to 5 would allow the retention of product **3** in the aqueous phase along with amine **2**. Adjustment of the pH to 7 and extraction with a suitable organic solvent, for example toluene or anisole, would then result in close-to-quantitative separation of product **3** from amine **2** and
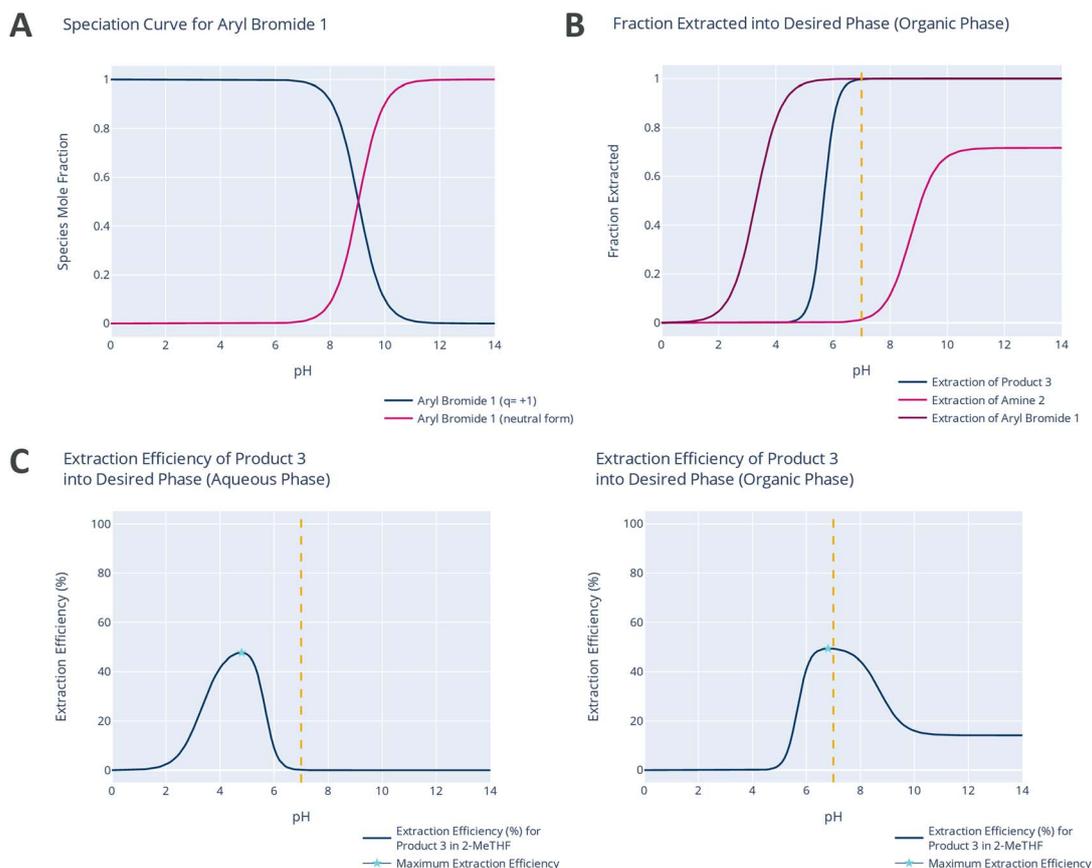
Fig. 3 Informative data visualisations for Scenario 2. Panel A: speciation curves of the aryl bromide **1**. Panel B: fraction extracted in the organic phase for all three components. Panel C: extraction efficiency of product **3** in either the aqueous (left) or the organic phase (right) in the presence of aryl bromide **1** and amine **2**. The pH value for the maximum extraction efficiency in each case is indicated with a star symbol. Volume ratio (organic/aqueous) = 1. Organic phase: 2-MeTHF. Vertical dashed line denotes pH 7.

its transfer in the organic phase. This sequence would maximise the recovery and residual purity of the isolated product **3**.

Both scenarios examined above represent common challenges that chemists face during design and optimisation of industrial processes. However, expertise to retrieve, format, or aggregate the necessary physical properties data and to model the liquid–liquid extraction step in a standardised manner is not common among process chemists. Our digital tool removes these data accessibility and expertise barriers and puts the data at the user's fingertips in a standardised layout. The visualisation of the speciation curves and fraction extracted in the respective phases facilitate process understanding by providing direct comparisons of the quantities examined.

Although the above functionalities have a positive impact on decreasing the effort required to model an aqueous liquid–liquid extraction process, the capability described so far is limited to a single, user-defined process solvent. We have designed the final part of the tool so that it can be used to compare the extraction efficiency of the desired compound in many systems, thus enabling the selection of solvents and process parameters. This calculation is fast, can be applied to multiple solvents, and provides the user with valuable information with which they can interact directly within the tool. In

both scenarios, the user has direct access to the data in tabular form (see ESI,† example of tool output), which can be sorted in ascending or descending order according to their criteria, for example by aqueous phase pH, solvent name, or extraction efficiency.

The tool can calculate the extraction efficiency of the desired product in the presence of an indeterminate number of impurities, in this case two, and 166 water-immiscible solvents over 140 pH increments (0.1-unit increments) in less than 5 seconds. In scenario 2, this calculation results in a cumulative data set of 22 701 entries. In this manner, the user can rapidly generate hypotheses regarding which solvents can be used in the liquid–liquid extraction step to maximise the extraction efficiency as a function of pH and solvent. In scenario 2, the user can compare between all 166 available solvents to identify a suitable option for the orthogonal extraction to separate product **3** from amine **2**. We highlighted toluene and anisole as examples of solvents used abundantly in industrial processes which also show excellent extraction efficiency for product **3** in a wide pH range (6–8), which de-risks the overall process (see Fig. 4).

It is important to point out the limitations of the current implementation of the tool, to manage user expectations and to ensure a correct use of the results. Currently, the LogP values
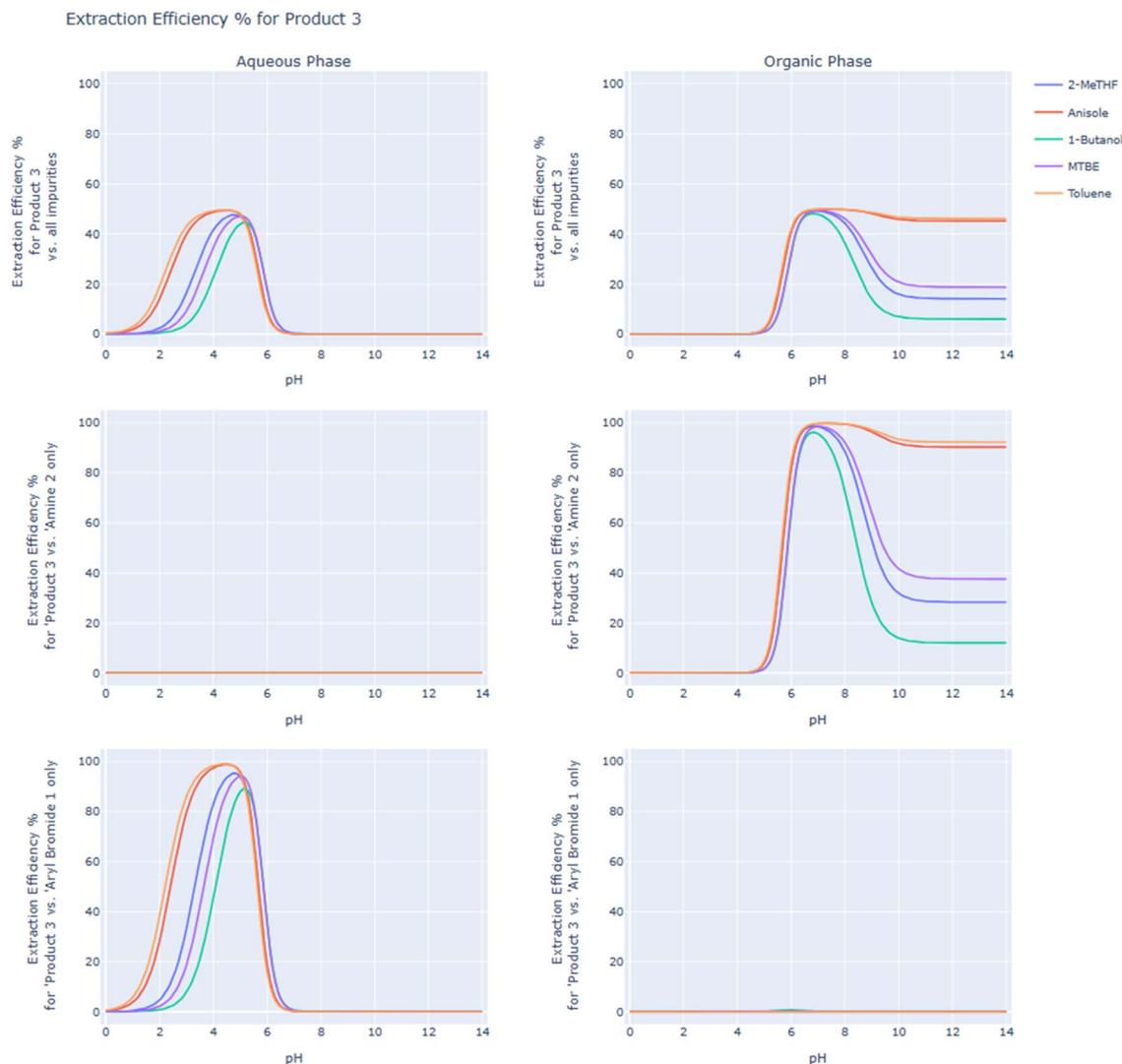
Fig. 4 Extraction efficiency curves of product **3** against both bromide **1** and amine **2** (top), or either amine **2** (middle) or bromide **1** (bottom). This visualisation highlights the opportunity of using orthogonal extractions to efficiently isolate product **3** by sequentially separating from the other two components of the mixture. The extraction efficiency in either the aqueous (left) or the organic (right) phase is shown. The calculation can be extended to many solvents highlighting potential opportunities. The plot is interactive allowing the visualisation of all selected solvents or just a subset of them. Included in this example: 2-MeTHF, 1-butanol, anisole, methyl *tert*-butyl ether, and toluene. Volume ratio (organic/aqueous) = 1.

used refer to 25 °C while many extractive processes can be conducted at a different temperature. Furthermore, due to the prediction approach taken, the stored LogP values may not be directly relevant to the actual process conditions.[15] The model computes partition coefficients at infinite dilution for binary aqueous/organic systems. Highly concentrated solutions, the presence of electrolytes in the aqueous phase or of ion pairs in the organic phase, as well as the presence of additional organic compounds (additional solutes or solvents) is not accounted for. Including such terms would require a more complex physical model that is not amenable to automation and suffers from a higher prediction error. As such, the tool is not geared to predict accurately a global optimum in the parameter space. Despite the limitations, the tool in its current implementation can be used for the virtual screening of solvents and to generate

hypotheses and prioritise them, guiding experimentation in an efficient and data-led manner.

Depending on data availability, the tool can be adjusted to use a range of available data sources. These can include measured or predicted LogP values at one or multiple temperature points, measured or predicted pK_a values, and measured or predicted solubility values. The latter can be used to calculate relative volumes for the organic phase and using the volume ratio, the relative volumes of the aqueous phase and thus the total relative volume that the process would require. This calculation can be used to estimate PMI values which can then be used as a selection criterion connecting this digital tool directly with sustainability strategy and targets. Finally, the equations could be modified to accommodate for the molar ratio between compounds instead, thus providing an estimate

of the residual purity of the isolatable compound with a certain degree of confidence.

The simplicity of this digital tool allows even the inexperienced user to readily set up an investigation involving an array of conditions including different solvents, volume ratios, and pH values to supplement existing experimental data or guide experimental design, leading to the development of novel, leaner, and more sustainable processes.[16–19] We estimate that adoption of this tool as part of the routine approach to solvent selection in process development activities will result in substantial productivity improvements and reduction in process development lead times. In addition to the lowered entry barrier, easy access to data from a curated database, without need for specialist assistance, and automatic population of all relevant fields, is estimated to save at least 30 minutes per every extraction being evaluated. This estimation is in comparison to previous ways of working based on templated excel spreadsheets requiring manual data retrieval and input. The time saved for systematic comparison of all solvent options available in a single synthetic stage of a process, exceeds one full working day. Additional and more substantial time savings are expected to be realised by directing experimentation to achieve pre-set development goals more swiftly, compared to a human-driven approach based on a personal experience and trial-and-error approach. Widespread adoption of this tool is expected to also bring additional benefits to the above-mentioned time savings, such as improved knowledge retention, higher quality documentation, and reduction in laboratory waste.

Conclusions

We have developed an accessible, easy-to-use, digital tool that simulates aqueous liquid–liquid extractions and can support data-led experimental design. This tool accommodates all the main considerations of a chemist taking on the design of an extractive process. It removes barriers and allows the chemist to interact with the data, building their process understanding dynamically through worked examples. As discussed in the two example scenarios, the tool allows the user to exercise their chemical expertise and make informed decisions, which facilitate the design and development of sustainable processes for the preparation of active pharmaceutical ingredients, or fine chemicals in general. In Scenario 1, the extraction efficiency graph suggested extractive conditions which were very similar to those reported for the preparation and extractive isolation of product **3**. Although data for Scenario 2 do not exist, the suggestion for orthogonal sequential extractions is sensible and could be verified by any chemist. We have experienced the immediate adoption of this tool by our colleagues, and we anticipate that similar tools could impact positively the ways of working of similar departments in other organisations.

Data availability

Data and processing scripts for this paper, including table of LogP values, table of pK_a values, table of compounds' SMILES, table of Solvent physical properties, table of results as an example of tool output, and an interactive notebook with use of the code, are available on Github through: **https://github.com/AstraZeneca/LLE_Digital_Tool** and DOI: **https://doi.org/10.5281/zenodo.15363857**.

Author contributions

S. T. conceived the idea for the digital tool and oversaw the project. G. K. led the code development. G. K., T. M., and E. H. E. F. contributed to the code. M. T. developed the physical properties data pipelines. G. K. wrote the manuscript. All authors commented and approved the final version of the manuscript.

Conflicts of interest

All authors are employees of and shareholders in AstraZeneca, who funded the study.

Acknowledgements

References

1 R. Sheldon, in *Green Chemistry in the Pharmaceutical Industry*, 2010, pp. 1–20, DOI: **10.1002/9783527629688.ch1**.

2 C. S. Slater, M. J. Savelski, W. A. Carole and D. J. C. Constable, in *Green Chemistry in the Pharmaceutical Industry*, 2010, pp. 49–82, DOI: **10.1002/9783527629688.ch3**.

3 C.-K. Chen and A. K. Singh, *Org. Process Res. Dev.*, 2001, **5**, 508–513.

4 M. Pietrasik, A. Wilbik and P. Grefen, *Dig. Chem. Eng.*, 2024, 100161.

5 A. E. Rubin, S. Tummala, D. A. Both, C. Wang and E. J. Delaney, *Chem. Rev.*, 2006, **106**, 2794–2810.

6 D. Ivanov, D. Alexandre and B. Sokolov, *Int. J. Prod. Res.*, 2019, **57**, 829–846.

7 J. Gmehling, M. Kleiber, B. Kolbe and J. Rarey, *Chemical thermodynamics for process simulation*, John Wiley & Sons, 2019.

8 A. I. Vogel, *A Textbook of Practical Organic Chemistry*, 1948.

9 I. W. Ashworth and R. E. Meadows, *J. Org. Chem.*, 2018, **83**, 4270–4274.

10 E. Müller, R. Berger, E. Blass, D. Sluyts and A. Pfennig, Thermodynamic Fundamentals, in *Ullmann's Encyclopedia of Industrial Chemistry*, 2008, ch. 2, DOI: **10.1002/14356007.b03_06.pub2**.

11 H.-J. Federsel, M. Hedberg, F. R. Qvarnström and W. Tian, *Org. Process Res. Dev.*, 2008, **12**, 512–521.

12 I. ICH, *ICH-Endorsed Guide for ICH Q*, 2011, vol. 8.

13 J. Yano, K. J. Gaffney, J. Gregoire, L. Hung, A. Ourmazd, J. Schrier, J. A. Sethian and F. M. Toma, *Nat. Rev. Chem.*, 2022, **6**, 357–370.

14 R. Mercado, S. M. Kearnes and C. W. Coley, *J. Chem. Inf. Model.*, 2023, **63**, 4253–4265.

15 C. Loschen, J. Reinisch and A. Klamt, *J. Comput.-Aided Mol. Des.*, 2020, **34**, 385–392.

16 J. R. Schmink, A. Bellomo and S. Berritt, *Aldrichimica Acta*, 2013, **46**, 71–80.

17 B. L. Y. S. Sanghvi, *Org. Process Res. Dev.*, 2015, **19**, 685–686.

18 B. Cohen, D. Lehnherr, M. Sezen-Edmonds, J. H. Forstater, M. O. Frederick, L. Deng, A. C. Ferretti, K. Harper and M. Diwan, *Chem. Eng. Res. Des.*, 2023, **192**, 622–637.

19 M. A. Ashley, M. H. Aukland, M. C. Bryan, M. A. Cismesia, T. Dutschei, O. D. Engl, P. S. Engl, Á. Enriquez Garcia, V. Harawa, G. Karageorgis, C. B. Kelly, A. Leclair, J. W. Lee, Z. Lei, W. Li, J. Pawlas, P. F. Richardson, S. C. Scott, A. Steven, B. S. Takale, D. Yerkozhanov and M. Zeng, *Org. Process Res. Dev.*, 2024, **28**, 3450–3459.