Systematic Optimization of Liquid–Liquid Extraction for Isolation of Unidentified Components

Sofja Tshepelevitsh,† Kertu Hernits,† Jaroslav Jenčo,‡ Joel M. Hawkins,§ Koji Muteki,§ Petr Solich,‡ and Ivo Leito*†

†Institute of Chemistry, University of Tartu, 14a Ravila Street, Tartu 50411, Estonia
‡Faculty of Pharmacy, Department of Analytical Chemistry, Charles University, Akademika Heyrovského 1203, Hradec Králové 50005, Czech Republic
§Pfizer Worldwide R&D, Eastern Point Road, Groton, Connecticut 06340, United States

ABSTRACT: We present a systematic approach for predicting the best solvents for selective extraction of components with unknown structure from complex mixtures (e.g., natural products)—a tool promising dramatic simplification of extraction process optimization. Its key advantage is that identification of the component(s) is unnecessary—prediction is based on a small set of experimental distribution coefficients (obtained using a combination of shake-flask extraction and chromatographic analysis) rather than structure-based descriptors. The methodology is suitable for the very common situations in practice where the desired compound needs to be separated from unknown impurities (i.e., selectively extracted from the mixture), as well as for large-scale and high-throughput work. The proof-of-concept methodology was developed and evaluated using an extensive set of experimental distribution data of lignin-related compounds obtained in this work.

INTRODUCTION

Liquid–liquid extraction is a separation method widely used both on laboratory scale and in industrial processes due to its simplicity, low cost, suitability for thermally labile and high-boiling compounds, and easy scalability. Properly chosen conditions and parameters (extraction solvents and/or extractants, pH, temperature) make it possible to concentrate the compounds of interest and separate them from undesirable components in a single step. However, systematic improvement of the process by predicting the effects of various parameter modifications is a challenging task. Factorial design of experiments is a useful, although comparatively labor-intensive, approach for optimization of complex conditions. Many computational methods (e.g., UNIFAC, linear free-energy relationship (LFER) models, and COSMO-RS) can replace experimental investigation in cases when the components of the mixture are known. However, systematic optimization of the process by analyzing a large number of potential solvent pairs is still, to the best of our knowledge, a problem without a solution. It is possible to estimate partition properties of a substance by its reverse-phase chromatographic retention time, yet practically useful predictions by this method are only expected for subsets of fairly similar solutes. Several methodologies for high-throughput determination of distribution coefficients (log D) are available, but further reducing the experimental workload would still be desirable.

It is known that partition of a neutral solute between water and an organic solvent may be satisfactorily described by a LFER equation employing five parameters of the solute: the descriptors of hydrogen bond (HB) acidity, HB basicity, dipolarity, polarizability, and molar volume. For ionizable solutes, at least one additional descriptor accounting for the properties of the ion is required. Obviously, the principle can be applied in reverse, and various properties of the solute can be deduced from the pattern of its partition/distribution coefficients in different solvent pairs. Thus, having established quantitative connections between the log D values of some model compounds in a variety of solvent pairs, one can use them to predict distribution properties of an unknown compound from a relatively small number of experimentally determined values. Unlike computations, experimental log D determination is possible without identification of the solute(s),

Received: September 28, 2017
Accepted: October 31, 2017
Published: November 10, 2017
even in the case of complex mixtures (see the Supporting Information for details and examples).

### Training Data and Prediction Algorithm

Log $D$ values for model compounds (training data) may be obtained experimentally and/or computationally, provided the computed values are of sufficient quality. In this work, we used full factorial experimental design leading to the full data matrix of experimental or experiment-derived $\log P / \log D$ values of 30 lignin-related compounds (depolymerization products and alike) in 20 solvent pairs. As lignin is one of the most abundant renewable materials on Earth and also a byproduct of several technological processes, its valorization is increasingly gaining interest.\(^{16,17}\) The solutes included 24 neutral or weakly acidic compounds (mostly phenolic) and 6 aromatic carboxylic acids; the solvent pairs were composed of 10 organic solvents of various polarity and hydrogen-bonding ability (Figure 1) in combination with 2 aqueous phases: neutral (pure water or phosphate buffer with pH = 7) and acidic (0.1 M HCl). Neutral aqueous phase allows the effective separation of carboxylic acids, which undergo extensive ionization at pH 7, from weaker acids and neutral compounds. A sufficiently basic (pH > 10) aqueous phase would likewise provide separation of weak acids from nonionizable compounds, but could not be included in this dataset due to experimental problems caused by alkaline solutions.

Various statistical methods may be used to build a model for structure-independent predictions, from the simple multilinear regression (MLR) used in this work to the more sophisticated principal component analysis.\(^{18}\) Here, we do not aim to establish a well-defined quantitative structure–activity relationship-like model, but rather to explore the potential of the created optimization strategy while allowing for adjustment according to the circumstances. The proposed principal steps for identification of the most effective/selective solvent pair are as follows:

1. adjust the training data (exclude unsuitable solvent pairs, if any, and compounds known/suspected to be considerably different from the solute of interest);
2. select $N$ reference solvent pairs in which the unknown solute(s) will be studied experimentally;
3. establish (multilinear) relationships between log $D$ in reference solvent pairs and the ones in which distribution is to be predicted (see eq 1);
4. determine log $D$ values in reference solvent pairs experimentally;
5. use equations from (c) and data from (d) to predict log $D$ values in all involved solvent pairs. Choose the one best suiting your purpose.

With the MLR approach used in this work, the mathematical model developed in step (c) takes the form of eq 1

$$\log D(X) = a_X + \sum_{i=1}^{N} (\rho_X(RSP_i) \cdot \log D(RSP_i))$$

where $X$ denotes a solvent pair in which the solute was not studied experimentally, RSP, is a reference solvent pair, $a_X$ and $\rho_X$ are the corresponding model parameters (one constant $a$ and $N$ coefficients $\rho$ for each $X$). The comprehensive details of the algorithm, program codes, and numerical examples of model parameters are provided in the Supporting Information.

### Results and Discussion

Complexity of the relationships between log $D$ values in different solvent pairs increases with the growing differences between the solvents, e.g., their varying abilities for $\pi-\pi$ and HB interactions (Figure 2). The greater the diversity of the solvent pairs in the training data, the more experimental values are needed for accurate predictions, so a compromise between effort and gain is to be found. The number of reference solvent pairs was optimized using leave-one-out cross validation. The results are presented in Table 1.

For comparison, an estimate of the day-to-day reproducibility of the experiments, expressed as the weighted pooled standard deviation $s$ (calculated by eq S3 in the Supporting Information) is 0.08 log units. The behavior of carboxylic acids was predicted less accurately than that of the weakly or nonacidic solutes, which could be caused by either under-representation of acids in the training data or underaverage quality of the corresponding experimental values ($s = 0.13$). In further work, we used five reference solvent pairs, which was reduced to four when the solute of interest appeared to be nonionizable.

To test the suitability of the model for predicting the most selective conditions, the solvent pairs providing the greatest differences of log $D$ values of the two solutes were predicted for all possible pairs of training set compounds. With 5 input log $D$ values and the full training set, the most selective solvent pairs were correctly identified in 60% of cases, predicted to be second-best in 24% of cases, and third-best in 6% of cases. Therefore, with high-quality data and dilute solutions, the actual best solvent pair is among the predicted 3 best solvent pairs in 90% of cases.

However, the created predictive approach relies on two assumptions that may not hold under real working conditions: (1) there are no interactions between the components of the mixture that could significantly affect their partition, e.g.,

---

**Figure 1.** Organic solvents and exemplary subset of solutes (see full list in the Supporting Information) included in the training data.
Acids, and abietic acid, respectively. Peaks of several main components of these substances are sesquiterpenes, fatty products: cedar oil, black cumin seed oil, and colophony. The points corresponds to the number of HB donor groups in the solute molecule; the size of the points reflects the specific volume of the solute (in Å³ amu⁻¹).

Real-life performance of the prediction algorithm was evaluated via experiments with “difficult” mixtures: relatively concentrated solutions (1–10%) of three plant-derived natural products: cedar oil, black cumin seed oil, and colophony. The main components of these substances are sesquiterpenes, fatty acids, and abietic acid, respectively. Peaks of several unidentified components allowing for sufficiently accurate integration and logD calculation were chosen on the chromatograms (see the Supporting Information for details).

As expected, the absolute errors of prediction are larger in the case of difficult mixtures than those in the case of training data: the overall RMSE (in log units) is 0.3 for cedar oil, 0.5 for cumin oil, and 0.4 for colophony. This may in part be caused by more chromatographic interferences resulting in less reliable experimental data. By far the largest prediction errors (Figure 3) belong to the solvent pairs including hexane—the least polar of the included solvents. An additional series of experiments (detailed in the Supporting Information) proved that extraction properties of low-polarity solvents when applied to polar solutes are strongly affected by the presence of relatively polar compounds (including the solute itself) in the organic phase (a phenomenon called the “matrix effect” here). At the same time, the extractive properties of HB-active solvents proved virtually free of such influences. If an as accurate as possible prediction is required for a solvent pair susceptible to matrix effects, one should include similarly susceptible solvent pair(s) in the model input. In the case of our data, inclusion of a matrix-effect-affected value into the model input did not dramatically change the quality of prediction for nonsusceptible solvent pairs.

Despite the relatively low absolute accuracy of prediction, the patterns of solute extraction behavior are reproduced well enough to be of help in solving practical problems. The most effective (i.e., providing the highest logD of the solute) extraction solvent of those not used as model input was identified correctly in 5 cases out of 11 and predicted to be second- or third-best in another 4 cases, which can be considered satisfactory in view of the relatively narrow spans of the observed logD values.

A series of experiments also showed that the applicability domain of the model may be easily expanded from single-component organic phases to certain binary solvents, and the effect of water-miscible cosolvent can be accounted for with relatively little effort. However, the accuracy of the correction will depend on the relative characteristics of the cosolvent and the primary solvent. See the Supporting Information for more details.

**CONCLUSIONS**

It was demonstrated that logD values of an unknown solute in 6–15 solvent pairs may be predicted using 4 or 5 experimental log D values with an error of 0.2–0.5 log units (depending on the magnitude of interferences), which is comparable to experimental errors of high-throughput log D determination. Even when lacking absolute accuracy, the predictions are useful for practical purposes, e.g., identification of the solvent pairs providing the best extraction efficiency or selectivity. The applicability domain of the approach can be potentially extended to the optimization of other parameters (e.g., aqueous phase modifiers, cosolvents, and temperature).
phase liquid-chromatographic analysis of the formed organic and aqueous phases with UV–vis detection. Most extractions were carried out with multisolute solutions containing usually 6–7 compounds, with a total concentration of circa 1.5 mg mL⁻¹ (total mass of solutes divided by the total volume of liquids). Suitable chromatographic methods were created to achieve the separation of the solutes from each other and from the UV–vis-active organic solvents. The experiments were carried out at an ambient temperature of 23 ± 2 °C.

More experimental details, origin of the chemicals and equipment, detailed evaluation of accuracy and error sources, and the obtained distribution coefficient data are given in the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.7b01445.

Experimental details, obtained log D data, assessment of data precision and accuracy, detailed computational algorithms with computer codes, additional experiments with mixed solvents (PDF)

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: ivo.leito@ut.ee. http://analytical.chem.ut.ee/.

ORCID
Ivo Leito: 0000-0002-3000-4964

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the EU through the European Regional Development Fund (TK141 “Advanced materials and high-technology devices for energy recuperation systems”) and by the Institutional funding project IUT20-14 from the Estonian Ministry of Education and Research. J.J. gratefully acknowledges support by the Charles University, Faculty of Pharmacy in Hradec Králové (Project SVV 260 412) and the support of the Erasmus+ program of the European Union.

■ REFERENCES

(1) Solvent Extraction Principles and Practice, Revised and Expanded; Rydberg, J., Cox, M., Musikas, C., Choppin, G., Eds.; CRC Press, 2004.
(2) Tedder, D. W. Liquid-Liquid Extraction. In Albright’s Chemical Engineering Handbook; CRC Press, 2008; pp 709–735.
(3) Jowkardeh, M.; Raoof, F. Optimization of Supercritical Fluid Extraction Combined with Dispersive Liquid–Liquid Microextraction as an Efficient Sample Preparation Method for Determination of 4-Nitrotoluene and 3-Nitrotoluene in a Complex Matrix. Talanta 2012, 88, 50–53.
(4) Ahmadi, K.; Abdollahzadeh, Y.; Asadollahzadeh, M.; Hemmati, A.; Tavakoli, H.; Torkaman, R. Chemometric Assisted Ultrasound Leaching-Solid Phase Extraction Followed by Dispersive-Solidification Liquid–Liquid Microextraction for Determination of Organophosphorus Pesticides in Soil Samples. Talanta 2015, 137, 167–173.
(5) Mylonaki, S.; Kiassos, E.; Makris, D. P.; Kefalas, P. Optimisation of the Extraction of Olive (Olea Europaea) Leaf Phenolics Using Water/Ethanol-Based Solvent Systems and Response Surface Methodology. Anal. Bioanal. Chem. 2008, 392, 977–985.
(6) Wienke, G.; Gmeining, J. Prediction of Octanol-Water Partition Coefficients, Henry Coefficients and Water Solubilities Using UNIFAC. Toxicol. Environ. Chem. 1998, 65, 57–86.

---

EXPERIMENTAL SECTION

The distribution ratios of ligneous compounds were determined using the shake-flask method followed by reverse-
(7) Abraham, M. H.; Acree, W. E. Equations for the Transfer of Neutral Molecules and Ionic Species from Water to Organic Phases. J. Org. Chem. 2010, 75, 1006−1015.

(8) Klamt, A. COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design; Elsevier Science Ltd: Amsterdam, 2005.

(9) Berthod, A.; Carda-Broch, S. Determination of Liquid−liquid Partition Coefficients by Separation Methods. J. Chromatogr. A 2004, 1037, 3−14.

(10) Poulsen, C. E.; Wootton, R. C. R.; Wolff, A.; deMello, A. J.; Elvira, K. S. A Microfluidic Platform for the Rapid Determination of Distribution Coefficients by Gravity-Assisted Droplet-Based Liquid−Liquid Extraction. Anal. Chem. 2015, 87, 6265−6270.

(11) Bao, J. J.; Liu, X.; Zhang, Y.; Li, Y. The Development of a High-Throughput Measurement Method of Octanol/Water Distribution Coefficient Based on Hollow Fiber Membrane Solvent Microextraction Technique. J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 2014, 907, 183−189.

(12) Abolhasani, M.; Coley, C. W.; Jensen, K. F. Multiphase Oscillatory Flow Strategy for in Situ Measurement and Screening of Partition Coefficients. Anal. Chem. 2015, 87, 11130−11136.

(13) Shalaeva, M.; Caron, G.; Abramov, Y. A.; O’Connell, T. N.; Plummer, M. S.; Yalamanchi, G.; Farley, K. A.; Goetz, G. H.; Philippe, L.; Shapiro, M. J. Integrating Intramolecular Hydrogen Bonding (IMHB) Considerations in Drug Discovery Using Δlog P As a Tool. J. Med. Chem. 2013, 56, 4870−4879.

(14) Yamagami, C.; Hamasaki, A.; Kumagai, Y.; Moritoki, T.; Tanahashi, T. Hydrogen-Bonding Abilities for Phenols Assessed by Quantitative Analyses of Their Partition Coefficients Derived from Different Partitioning Systems. Chem. Pharm. Bull. 2005, 53, 398−401.

(15) Zissimos, A. M.; Abraham, M. H.; Barker, M. C.; Box, K. J.; Tam, K. Y. Calculation of Abraham Descriptors from Solvent−water Partition Coefficients in Four Different Systems; Evaluation of Different Methods of Calculation. J. Chem. Soc., Perkin Trans. 2 2002, 470−477.

(16) Rinaldi, R.; Jastrzebski, R.; Clough, M. T.; Ralph, J.; Kennema, M.; Bruynincx, P. C. A.; Weckhuysen, B. M. Paving the Way for Lignin Valorisation: Recent Advances in Bioengineering, Biorefining and Catalysis. Angew. Chem., Int. Ed. 2016, 55, 8164−8215.

(17) Zakzeski, J.; Bruynincx, P. C. A.; Jongerius, A. L.; Weckhuysen, B. M. The Catalytic Valorization of Lignin for the Production of Renewable Chemicals. Chem. Rev. 2010, 110, 3552−3599.

(18) Kettaneh-Wold, N.; MacGregor, J. F.; Dayal, B.; Wold, S. Multivariate Design of Process Experiments (M-DOPE). Chemom. Intell. Lab. Syst. 1994, 23, 39−50.

(19) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. Chemom. Intell. Lab. Syst. 1987, 2, 37−52.