Correlating the partitioning of organic molecules between water and [MeoeMPyrr]+ [FAP]- through machine learning

Flora Esposito, a Ulderico Di Caprio, a Florence Vermeire, b Mumin Enis Leblebicia,\*

a Center for Industrial Process Technology, Department of Chemical Engineering, KU Leuven, Agoralaan Building B, 3590 Diepenbeek, Belgium

b KU Leuven, Department of Chemical Engineering, Celestijnenlaan 200F-bus 2424, Leuven 3001, Belgium

\*muminenis.leblebici@kuleuven.be

Abstract

Lipophilicity is one of many parameters involved in the biological activity of drugs. It is assessed by defining the partitioning of a molecule between an organic (i.e. octanol) and a water phase. Nevertheless, octanol is too simple to encode all of the complicated interactions seen in ionic liquids. Moreover, the experimental determination of logP in specific ionic liquid/water systems (logPIL/W) is an arduous and resource-intensive task. Machine learning and hybrid modelling techniques have emerged as essential tools in chemical engineering, providing innovative solutions to complicated physicochemical problems. This study proposes a hybrid model correlating the partitioning of organic molecules in octanol/water with the partitioning in [MeoeMPyrr]+[FAP]-/water systems. The hybrid model is formed by a first principle and a data-driven part. The first is represented by a group contribution model, the latter is represented by an ensemble of 5 Multilayer Perceptron (MLP) models. The model structure with the highest accuracy and generalization properties is searched with 5-fold cross-validation, using hyperparameter optimization.. The prediction capabilities have been evaluated through various metrics, namely the coefficient of determination (R2 = 0.93), the Mean Squared Error (MSE = 1.8·10-2) and the Mean Absolute Percentage Error (MAPE = 25%), showing quite good accuracy.

**Keywords**: Hybrid Modelling; Octanol/Water Partitioning Coefficient; Artificial Neural Network; Ensembling; Machine Learning

* 1. Introduction

The octanol/water partition coefficient (logP) plays a pivotal role in drug discovery, environmental chemistry, and industrial processes (Kujawski et al., 2012). It is defined as the partitioning of a molecule between an organic (i.e., octanol) phase and a water phase. As octanol is too simple to encode all of the interactions in more complicated systems, manufactured and natural membranes, such as liposomes and micelles, have lately been exploited as nonaqueous phase alternatives (Loureiro et al., 2018). An ionic liquid (IL) is a salt that is liquid at room temperature. ILs find diverse commercial applications, such as chemical synthesis solvents, gas chromatography stationary phases, liquid matrices for compound isolation, and absorbents (Jiang et al., 2013). Experimental measurements have also shown their potential as solvents for CO2 absorption. The partitioning mechanism with water is critical for understanding the movement of analytes from the aqueous to the IL phase, which happens during the extraction of hydrocarbons and phenols. However, the available partition data between ILs and water for organic compounds is very limited (Padró et al., 2011). Moreover, the experimental determination of logPIL/W in specific ionic liquid/water systems is an arduous and resource-intensive task. Lipophilicity has traditionally been measured using methods such as water-octanol shake flask and high-performance liquid chromatography. These procedures are time-consuming and can require multistep dilutions (Wenlock et al., 2011).Chemical engineers rely on models for design, research, and day-to-day decision-making. Many recent initiatives have played a role in the spread of machine learning techniques in the research field by creating extensive databases, benchmarks, and representations for chemical applications (Heid et al., 2023). Chemical systems, on the other hand, should adhere to physical principles such as conservation laws and constitutive equations. Hybrid modelling is a methodology that benefits from the synergistic use of first principle and data-driven models, by combining existing process knowledge and information provided by collected data. Such a strategy is especially appropriate for systems and sectors where data creation requires a large amount of resources (Narayanan et al., 2023). This paper proposes a hybrid model structure capable of predicting the logPIL/W in the [MeoeMPyrr]+[FAP]-/water system.

* 1. Dataset description and preprocessing

The data to train the model were curated from literature (Jiang et al., 2013). The authors provide the experimental partition coefficient of organic molecules in [MeoeMPyrr]+ [FAP]- I/Water systems. Simple organic molecules are present in the dataset, such as Acetophenone, Aniline, Benzene and so on. The dataset used consists of 41 experimental points. It contains the names, Chromatographic retention factors (logK) and logPIL/W of organic molecules in [MeoeMPyrr]+ [FAP]- I/Water systems. In this work, only the experimental logPIL/W have been employed. Three points were excluded from the dataset since the ratio of logP in the two different solution systems is much lower than one (i.e , , ). Those values are outside the range of expected ones. To robustly evaluate the model performances, a cross-validation strategy is applied. From the original dataset, two sets are randomly extracted: the training and the test sets. The training set contains 30 points and is used to train the network. The test set has 8 points. The MSE on the test set is calculated by k-fold cross-validation. Prior to the model training, the data were normalized in the range [0,1] using a min/max scaler.

* 1. Model description

The present section provides the main characteristics of the model used in this work. Correspondingly, we here describe the topology of the first principle and the ensemble of data-driven models composing the hybrid model.

* + 1. Hybrid Model Structure

Figure 1 depicts a brief representation of the model. It requires as input the Simplified Molecular Input Line Entry System (SMILES) of the chemical component and returns the expected value of the logPIL/W in the [MeoeMPyrr]+ [FAP]- /Water systems system. In the model pipeline, a molecular fragmented description is obtained from the SMILES, through RDKit library in Python. There are 68 fundamental atom types included with one-hot encoding. It includes metals and noble gases as well as the elements usually present in organic compounds (C, H, N, O, S, P, halogens). It is constructed such that each atom in the molecule matches just one type of atom, reducing ambiguity from the typing system (Wildman & Crippen, 1999). The descriptors are then employed as input features for an ensemble of 5 Multi-Layer Perceptron (MLP) models. The ensemble is an estimator that fits a number of regressors and then averages their outcomes to give a final prediction. This technique is used to estimate a proportionality coefficients () between logP values in the octanol/water system and those in the [MeoeMPyrr]+[FAP]–/water system that depends on the molecular structure of the solute.

|  |  |  |
| --- | --- | --- |
| . |  | (1) |

The logP values in the octanol/water system are found from the molecular structure applying the Crippen method (Wildman & Crippen, 1999). This is done using RDKit library in Python.

A diagram of a diagram

Description automatically generated

Figure 1: Block diagram of the hybrid model. In the figure, is the octanol/water partition coefficient, is the [MeoeMPyrr]+[FAP]— /water partition coefficient and is the proportionality coefficient between the two.

* + 1. Multilayer perceptron

The ensemble presented in this work is formed by 5 multilayer perceptrons (MLPs). They have been trained through the backpropagation algorithm and with no activation function in the output layer. In the other layers, the activation function used is the rectified linear activation unit (ReLu) The loss function is represented by the squared error. In order to avoid overfitting, an L2 regularization term has been added to the loss function. The strength of the regularization term has been chosen through a k-fold cross-validation. For all the 5 different MLPs, the hyperparameters can be found in Table 1.

* + 1. Ensemble of multilayer perceptrons

In machine learning and statistics, ensemble methods are widely used. Ensemble methods provide techniques for combining multiple single predictors to form a committee, resulting in better and more accurate results than a single predictor (Guzman et al., 2016). Training of the ensemble involves independently training multiple MLPs regressors with the same dataset, but not with the same initial model parameters. The expected output is then obtained by averaging all individual single regressors output. In other words, the final prediction of the ensemble for a given data point is the average of the predictions made by each individual model. Table 1 shows the structure of the 5 different MLPs used in the voting ensemble. These structures have been found through a k-fold cross-validation and a grid search, better detailed in Paragraph 4.

Table 1: Structure of the 5 MLPs used in the ensemble.

|  |  |  |  |
| --- | --- | --- | --- |
| MLP | Number of hidden layers | Neurons per layer | Regularization term |
| 1 | 1 | 6 | 1∙10-5 |
| 2 | 3 | 50,50,50 | 1∙10-2 |
| 3 | 1 | 38 | 1∙10-3 |
| 4 | 3 | 20,20,20 | 1∙10-2 |
| 5 | 2 | 20,20 | 1∙10-2 |

* 1. K-fold cross-validation

In statistical modelling, cross-validation is a standard practice. To avoid overfitting the training data, the model interpolation accuracy and generalization capabilities must be estimated (Di Caprio et al., 2023) The hyperparameters of the ML models have been determined using 5-fold cross-validation and a grid search on the training set. The entire dataset is first randomly shuffled, guaranteeing that the order of data points is randomized. The randomized dataset is then split into 5 folds. The data is shuffled again before each split. Each iteration of the k-fold cross-validation loop uses k-1 folds for training and the remaining fold for testing. This operation is done 5 times, with each fold only acting as the validation set once. Additionally, hyperparameter tuning is performed using a grid search. The grid search module in scikit-learn library in Python allows the definition of an estimator (i.e. the ensemble), a grid of parameters (i.e. a grid for the number of hidden layers, neurons per layer and for the strength of the L2 regularization term), and the chosen number of cross-validation folds. The models are trained and evaluated within the cross-validation loop, optimizing the model based on the mean squared error (MSE) on the test set as the evaluation metric. Following that, the model is trained using the whole training set. The model generalization capabilities are determined using the test set, formed by the remaining 8 data points not included in the training set.

* 1. Model Evaluation

The model accuracy and generalization performances are evaluated using various statistical criteria, namely the coefficient of determination (*R2*), Mean squared error (MSE), and mean absolute percentage error (MAPE) for both the training and test sets.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |
|  |  | (3) |
|  |  | (4) |

where Ndataset is the sum of data points contained in the train and test sets, yi is the experimental output value of the ith point in the dataset, is the output value predicted at the ith point in the dataset, ymean is the average of the output values contained in the dataset.

* 1. Results

A graph of a graph with numbers and a line

Description automatically generated with medium confidenceA graph of a train set size

Description automatically generatedFigure 2 shows the prediction capabilities of the model structure with the highest accuracy on the validation set identified by the k-fold cross-validation. In this figure all the points are clustered around the ideal prediction line with a narrow dispersion. With few exceptions, both the predictions on the training and validation points indicate errors of less than 20%.

Figure 3: Mean square error (MSE) and mean absolute percentage error (MAPE) on the train and test sets as a function of the train set size.

Figure 2: Prediction plot of the model with the highest accuracy on the validation set.

The most challenging points to model are, as expected, the extreme values of the dataset. As we can see from the plot, the extreme values are indeed the ones with the highest error. Those points represent the partitioning of 1-Butanol and 1-Nitropropane. The R2 on the train and test sets is 0.93 and the MSE is 1.8·10-2. The metrics on the test set only are, instead: R2 = 0.87, MSE = 2·10-2 and MAPE = 15%. This highlights accuracy and generalization capabilities of the most performant ensemble identified by the 5-fold cross-validation. The voting ensemble average MSE is 2.29 ·10-2, slightly higher than the one obtained with the best MLPs structures. Following, the influence of the size of the training set on the model accuracy was investigated. As stated before, experimental measurements for specific ILs/water partitioning can be costly. Therefore, it could be useful to provide a model that can work with a relatively low number of experimental points. Figure 3 shows the results of the model trained with a decreasing number of points. Both the MSE and the MAPE are calculated on the train and test sets. From this figure, an increasing MSE and MAPE can be appreciated with a decreasing number of points, as expected. On the other hand, the lowest MSE is found with 28 training points with a value of 1.7·10-2. Looking at Figure 3, the MAPE seems to be high. Nevertheless, the values of range from -0.5 to 2.5. In particular, we can also notice a point for which = 0.006. Looking at (3), in the MAPE formula, the expected value is at the denominator. This usually increases the MAPE, if values close to 0 are taken into account. By excluding the aforementioned point from the metrics evaluation, the MAPE decreases to 15%.

* 1. Conclusions

A hybrid model for the prediction of logPIL/W in [MeoeMPyrr]+[FAP]— /water systems has been proposed. The model structure returns accurate predictions with an R2= 0.93 and good generalization capabilities. The best model structure is found through a k-fold cross-validation. Although the model developed in this paper shows acceptable prediction accuracy, its performance can be further improved. In particular, the first principle part of the model could be improved, providing more physical adherence. Moreover, the model can also be tested with a new set of experimental data in a different kind of organic-aqueous system. However, the hybrid model developed could be used to further analyze the partitioning of organic molecules in ionic liquid/water systems by performing a low amount of experiments.

* 1. References

Bannan, C. C., Calabró, G., Kyu, D. Y., & Mobley, D. L. (2016). Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *Journal of Chemical Theory and Computation*, *12*(8), 4015–4024.

Di Caprio, U., Wu, M., Vermeire, F., Van Gerven, T., Hellinckx, P., Waldherr, S., Kayahan, E., & Leblebici, M. E. (2023). Predicting overall mass transfer coefficients of CO2 capture into monoethanolamine in spray columns with hybrid machine learning. *Journal of CO2 Utilization*, *70*.

Guzman, E., El-Haliby, M., & Bruegge, B. (2016). Ensemble methods for app review classification: An approach for software evolution. *Proceedings - 2015 30th IEEE/ACM International Conference on Automated Software Engineering, ASE 2015*, 771–776.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., & McGill, C. J. (2023). *Chemprop: A Machine Learning Package for Chemical Property Prediction*.

Jiang, R., Anderson, J. L., Stephens, T. W., Acree, W. E. (William E., & Abraham, M. H. (Michael H. ). (2013). Abraham model correlations for predicting gas-to-liquid partition coefficients and activity coefficients of organic solutes dissolved in 1-(2-methoxyethyl)-1-methylpyrrolidinium tris(pentafluoroethyl) trifluorophosphate. *European Chemical Bulletin, 2013, Hungary: Deuton-X Ltd.*, *2*(10), 741–751.

Kujawski, J., Popielarska, H., Myka, A., Drabińska, B., & Bernard, M. (2012). The log P Parameter as a Molecular Descriptor in the Computer-aided Drug Design – an Overview. *Computational Methods in Science and Technology*, *18*(2), 81–88.

Loureiro, D. R. P., Soares, J. X., Lopes, D., Macedo, T., Yordanova, D., Jakobtorweihen, S., Nunes, C., Reis, S., Pinto, M. M. M., & Afonso, C. M. M. (2018). *Accessing lipophilicity of drugs with biomimetic models: A comparative study using liposomes and micelles*.

Narayanan, H., von Stosch, M., Feidl, F., Sokolov, M., Morbidelli, M., & Butté, A. (2023). Hybrid modeling for biopharmaceutical processes: advantages, opportunities, and implementation. *Frontiers in Chemical Engineering*, *5*, 1157889.

Padró, J. M., Ponzinibbio, A., Mesa, L. B. A., & Reta, M. (2011). Predicting the partitioning of biological compounds between room-temperature ionic liquids and water by means of the solvation-parameter model. *Analytical and Bioanalytical Chemistry*, *399*(8), 2807–2820.

Wenlock, M. C., Potter, T., Barton, P., & Austin, R. P. (2011). A method for measuring the lipophilicity of compounds in mixtures of 10. *Journal of Biomolecular Screening*, *16*(3), 348–355.

Wildman, S. A., & Crippen, G. M. (1999). Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, *39*(5), 868–873.