Solubility prediction of lipid compounds using machine learning

G. Gutiérrez Álvarez[a#], A. Porley Santana[a#], S. Gutiérrez Parodi[a] and J. Ferreira[a,b*]

[a] Grupo de Ingeniería de Sistemas Químicos y de Procesos, Instituto de Ingeniería Química, Facultad de Ingeniería, Universidad de la República, Montevideo, 11300, Uruguay
[b] Heterogenoeus Computing Laboratory, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, 11300, Uruguay
* Corresponding Author: jimenaf@fing.edu.uy
#These authors contributed equally to this work

ABSTRACT

Lipid purification processes are essential in lipid biomass valorization. Solubility is a key property in the solvent selection and process design. This work focuses on developing a predictive solubility model using machine learning techniques to optimize the separation of valuable compounds from a natural matrix derived from lanolin fat. First, the database was created from a literature review, then a database pre-processing step was performed, and the final step was model validation. Random Forest regression was selected for its ability to handle complex nonlinear relationships, showing better performance than bibliography models. An accurate model for lipids solubility in solvents was developed using machine learning techniques and experimental data.

Keywords: Solubility, Machine learning, Data preprocessing, Random Forest

INTRODUCTION

Biomass valorization not only reduces the environmental impact generated but also emerges as an opportunity for developing countries like Uruguay, which are agro exporters of high-quality products with a large amount of organic waste. In this context, wool grease, a complex mixture of organic compounds including esters of steroid and aliphatic alcohols, fatty acids, cholesterol, and other sterols is a by-product of wool washing [1]. To obtain high-value products, purification processes such as crystallization, chromatography, liquid-liquid extraction, among others are required [2,3]. In these processes, the interaction of the selected compound with a liquid phase known as a solvent or diluent (depending on the case) is a crucial aspect for the selection of the correct solvent in the design step [4]. Identifying the solvent solubility of the selected compounds is essential to achieve an effective separation.

Although, the determination of the solubility can be performed by several experimental methods, these methods are often labor-intensive and time consuming to reach the necessary equilibrium [5,6]. So, predictive solubility models are a powerful tool to accelerate solvent selection processes in process design. Solubility prediction models have evolved significantly over time. Hildebrand introduced the concept of the solubility parameter, providing a tool to predict solubility based on the thermodynamic properties of solvents and solutes [7]. Abraham later introduced a model that made use of a solubility parameters set, incorporating molecular descriptors that enable more precise modeling of specific solute-solvent interactions through linear free energy relationships [8]. Nowadays, a hybrid model has been developed that combines group contribution methods with machine learning (ML) algorithms to further improve the prediction of properties such as free energy and solvation enthalpy [9]. No specific studies have been found on predicting solubility in solvents for lipid compounds.

This research focuses on developing a machine learning-based predictive model to estimate the solubility of five key wool grease lipids (cholesterol, palmitic acid, oleic acid, stearic acid, and linolenic acid) along with naphthalene in a variety of organic solvents. The model is trained using experimentally derived data sourced from scientific articles and handbooks.

METHODOLOGY

This study began with a literature review, followed by dataset preprocessing. Subsequently, models were developed and validated.

The dataset was built by collecting data from 15 scientific papers, including [10-13], resulting in 728 data points for

6 organic solutes and 32 solvents. An analysis of existing solubility models, such as [8,9], was conducted to identify the properties to be used as input variables in solubility prediction. Twenty-one input variables were used to build the database. These include temperature, dipolar moment of solute and solvent, and Abraham parameters. Molecular weight and density were also included for comprehensive characterization. Database preprocessing only included unit homogenization, variable normalization, and correlation analysis (using Pearson and Spearman correlations). A 0.6 threshold was used to classify significant correlations, revealing several strongly correlated input variables.

The dataset was randomly split into 70% for training and 30% for testing. Three input variations were assessed: (S) a simple model with 9 uncorrelated variables, (A) a model that excludes Abraham parameters (16 variables), and (C) a comprehensive model that utilizes all 21 variables.

We used RF to build the three variations of the model and compare these models with the ones obtained by LR. The model search was conducted with the RF and LR implementations in Python via the Scikit-learn package. Hyperparameter optimization for RF was performed using Optuna [14], considering all algorithm-specific hyperparameters.

The performance and predictive ability of the models were evaluated in the validation set using the Mean Squared Error (MSE) and the Coefficient of Determination (R^2). Then, the model was compared to the Abrahams model (classic model) [8] and a hybrid model (based on group contribution methods and ML) [9].

RESULTS

Table 1 presents the values of RMSE and R^2 for the 3 variations of the input variables. In all cases. RF models perform better than the obtained by LR. Among the RF models, the most accurate was the variation A.

Table 1: Comparison of the (R^2) and the (MSE) between LR and RF models.

	LR		RF	
Variation	R^2	MSE	R^2	MSE
S	0.3035	0.0067	0.8276	0.0017
A	0.4529	0.0053	**0.9686**	**0.0003**
C	0.4679	0.0051	0.9435	0.0005

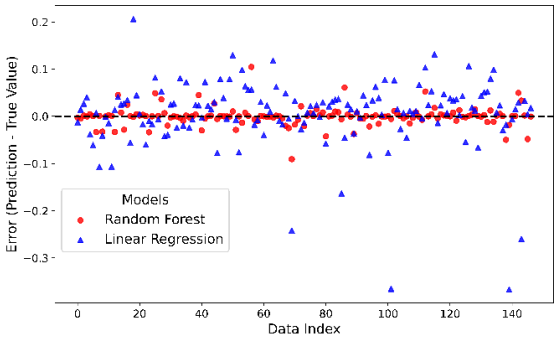

Figure 1: Comparative error predition of RF model and LR model (variation A).

As it's shown in Figure 1, RF demonstrates less error, indicating that it captures the nonlinearities of solubility more effectively than linear regression (LR) models. No significant differences were observed in the predictions for each compound.

Figure 2 shows the comparison of the RF model (variation A) with two state-of-the-art solubility prediction models (Abraham model [8] and a Hybrid model [9]). Concentric circles represent logarithmic increments in prediction error, with the center indicating perfect agreement. RF outperforms both state-of-art models, since they are generic models for organic compounds.

Although only variation A is shown in both figures, the behavior was similar for variations B and C.

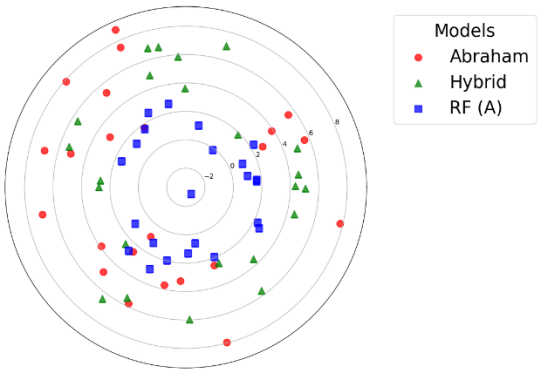

Figure 2: Model comparison at 298 K via log10($|y_{pred}–y_{real}|$) for solute–solvent combos (g/L).

CONCLUSIONS

The RF model enables more accurate predictions of organic compound solubility within the available data ranges than the linear regression model. All variants of the RF model exhibit better performance compared to the Abraham model and the hybrid model. Future efforts will involve extending the dataset to enhance prediction accuracy and robustness. Other ML methods with strong extrapolation capabilities will be explored. Furthermore, this model will be used as a basis for designing biomass lipid separation processes.

REFERENCES

1. Gutiérrez, S., Viñas M. (2003). *Water Sci. and Tech.* https://doi.org/10.2166/wst.2003.0379
2. Ding, H., et al. (2017). Chem. Pap. https://doi.org/10.1007/s11696-016-0043-1
3. Pei, H., et al. (2019). *J. of Separation Sci.* https://doi.org/10.1002/jssc.201900063
4. Diorazio, L. J., et al. (2016). Organic Proc. Res. & Dev. https://doi.org/10.1021/acs.oprd.6b00015
5. Atkins, P., de Paula J. (2014). *Physical Chemistry.* 10th edition.
6. Hermans, A., et al. (2022). *The AAPS J.* https://doi.org/10.1208/s12248-022-00760-8
7. Hildebrand, J. H. (1936). *Solubility of Non-Electrolytes.* 2nd edition.
8. Abraham, M. H., et al. (1987). *J. of Chromatography A.*

https://doi.org/10.1016/S0021-9673(01)86779-0
9. Chung, Y., et al. (2022). *J. of Chem. Info.and Modeling*. https://doi.org/10.1021/acs.jcim.1c01103
10. Flynn, G. L., et al. (1979). *J. of pharmaceutical sci.* https://doi.org/10.1002/jps.2600680908
11. Xu, Y.-S., et al. (2019). *J. of Chem. & Eng. Data*. https://doi.org/10.1021/acs.jced.9b00064
12. Calvo, B., Cepeda, E.A. (2008). *J. of Chem. & Eng. Data*. https://doi.org/10.1021/je7006567
13. Hoerr, C. W., Harwood, H.J. (1952). https://doi.org/10.1021/j150501a008
14. Akiba, T., et al. (2019). *25th ACM SIGKDD international conference*. https://doi.org/10.1145/3292500.3330701