Solubility prediction of lipid compounds using machine learning

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

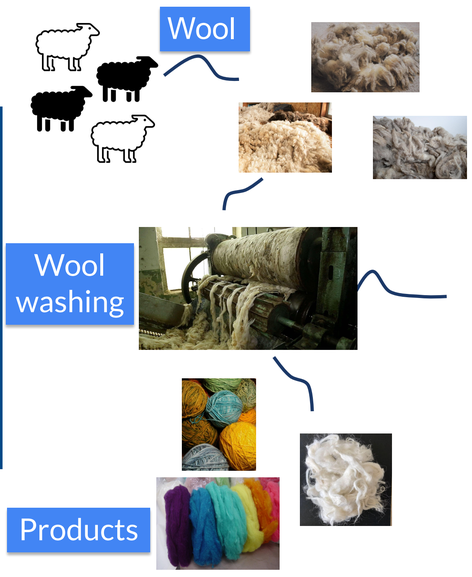
[a] Chemical & Process Systems Engineering Group, Faculty of Engineering, Universidad de la República, Montevideo, Uruguay

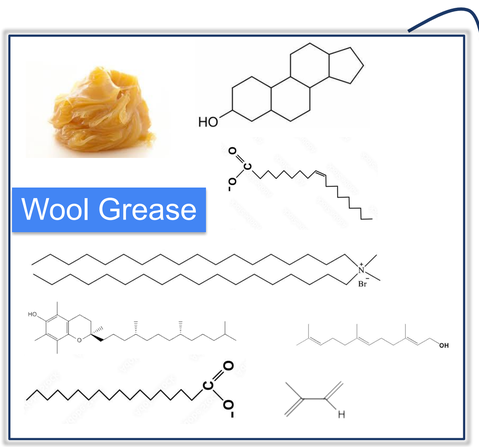
[b] Heterogeneous Computing Laboratory, Faculty of Engineering, Universidad de la República, Montevideo, Uruguay

* jimenaf@fing.edu.uy

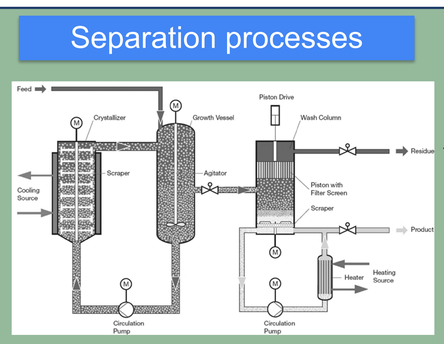
Introduction

Wool Processing





High-value products



- Crystallization
- Liquid Chromatography
- Liquid – Liquid extraction

- Skin products
- Eye drops

Key property in the design of these processes.: Solubility

Objectives

- 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.
- Model is trained using experimentally derived data sourced from scientific articles and handbooks.

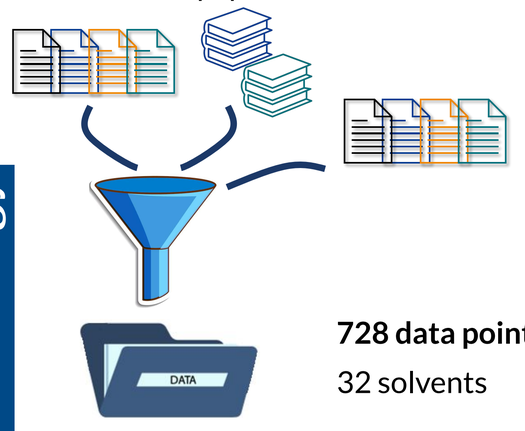
Solubility

- Solubility can be performed by several experimental methods: labor-intensive and time-consuming.
- Solubility models: powerful tool to accelerate solvent selection processes in process design.
- Not specific for lipids:
  - Model based on thermodynamic properties: Hildebrand [7].
  - Model based on thermodynamic and molecular descriptor: Abraham [8].
  - Hybrid model based on free energy and solvation enthalpy using ANN [9].

Methodology

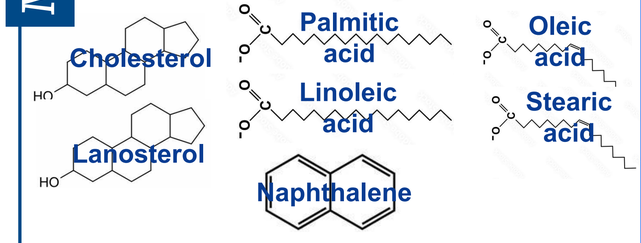
Data Collection

15 scientific papers and handbooks



728 data points

32 solvents



Data Preprocessing

- **21 input** variables
- Variable **normalization**
- Variable **selection**:
  - Pearson correlation
  - Spearman correlation

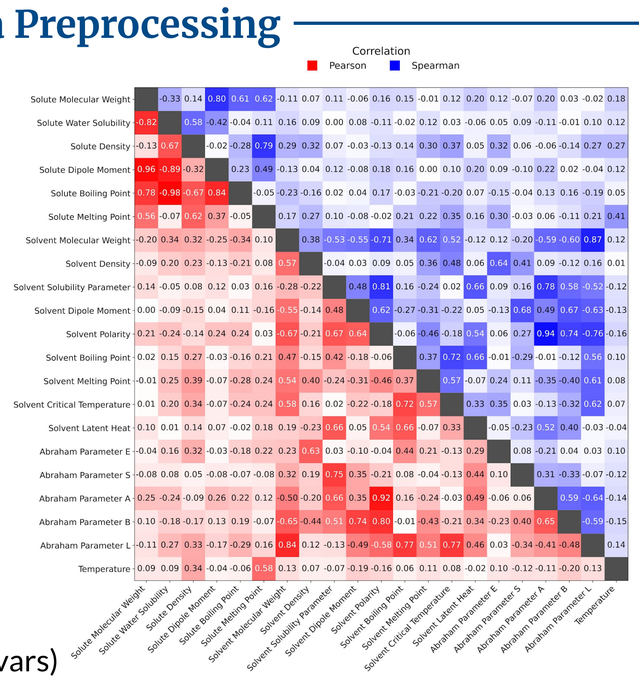
Temperature
Solute properties
Solvent properties
Abraham's solvation parameter

Randomly split

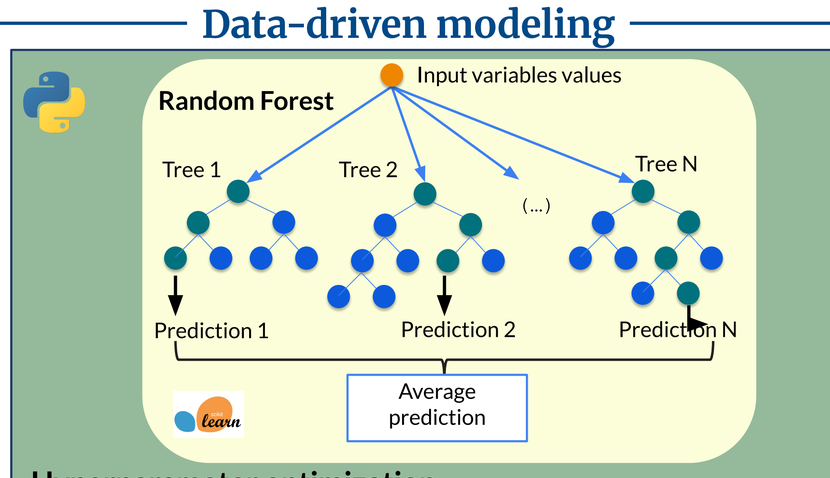
- Training: 70 %
- Testing: 30 %

Variations:

- Simple, S (9 uncorrelated vars)
- Without Abraham parameters, A, (16 vars)
- Complete (21 vars)



Data-driven modeling



Hyperparameter optimization

n_estimators ∈ [50 , 300]
max_depth ∈ [3 , 30]
min_samples_split ∈ [2, 10]
min_samples_leaf ∈ [1, 4]
max_features ∈ 'sqrt', 'log2', 0.1, 0.2, 0.5, 0.8
bootstrap ∈ True, False

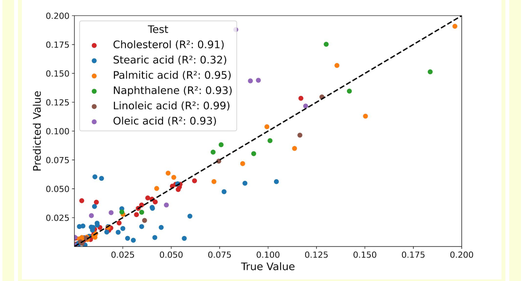
OPTUNA

Minimize: MSE
n_trials = 3000 / exploration_percentage = 0.5

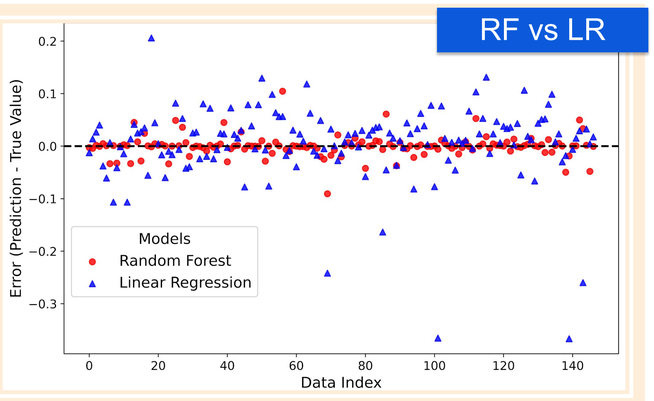
Results and discussion

	Linear Regression		Random Forest	
Variation	R^2	MSE	R^2	MSE
S	0.30	6.7×10^{-3}	0.83	1.7×10^{-3}
A	0.45	5.3×10^{-3}	**0.97**	**0.3×10^{-3}**
C	0.47	5.1×10^{-3}	0.94	0.5×10^{-3}

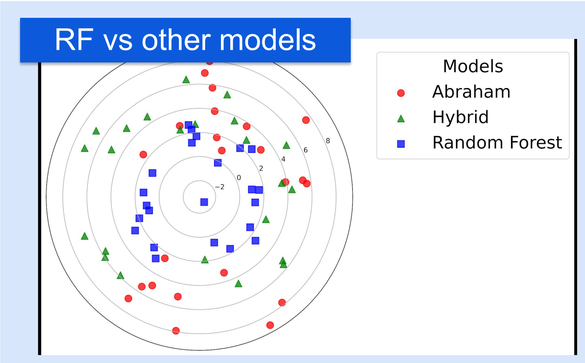
Prediction by solute



RF vs LR



RF vs other models



Future work

- Extend the dataset to enhance prediction accuracy and robustness.
- Other machine learning methods with strong extrapolation.
- The obtained 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

Acknowledgement CSIC

Instituto de Computación
Instituto de Ingeniería Química
Facultad de Ingeniería
Universidad de la República Uruguay