# EMJM Chemoinformatics+ Summer School 2025

Ljubljana, Slovenia — June 30 – July 4, 2025

## Predicting Aqueous Solubility of Drug-like Molecules Using Machine Learning

**Alice Smith¹, Bob Johnson², Carol Lee³**¹Department of Chemistry, University of Example, London, UK
²Institute of Computational Sciences, DataVille, Germany
³Faculty of Pharmacy, University of Ljubljana, Ljubljana, Slovenia
Corresponding Author Email: alice.smith@example.edu

## Abstract

Accurate prediction of aqueous solubility is a critical task in the early stages of drug development. In this study, we evaluate several machine learning approaches for predicting solubility using a curated dataset of drug-like molecules. Features were generated using molecular fingerprints and physicochemical descriptors. Random Forest, Support Vector Regression, and Gradient Boosting models were trained and compared using cross-validation. Our results show that ensemble methods outperform traditional models, achieving an RMSE of 0.45 logS units on the test set. This work demonstrates the potential of data-driven models in supporting medicinal chemistry workflows and provides a foundation for integrating predictive analytics into cheminformatics pipelines. As shown in **Figure 1**, the model architecture provides clear insight into feature importance.

[Insert Figure 1 Here]

Figure 1. Architecture of the machine learning model used for solubility prediction.

**Keywords:** aqueous solubility; machine learning; cheminformatics; drug design; predictive modeling

## References

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. \*Drug Discovery Today, 23\*(6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039

Lipinski, C. A. (2004). Lead- and drug-like compounds: The rule-of-five revolution. \*Drug Discovery Today: Technologies, 1\*(4), 337–341. https://doi.org/10.1016/j.ddtec.2004.11.007

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. \*Journal of Chemical Information and Modeling, 50\*(5), 742–754. https://doi.org/10.1021/ci100050t