Chapter 18
Modeling Log S:
Evaluating a Novel Predictive Approach Using Machine Learning on Diverse Datasets

Imane Aitouhanni
https://orcid.org/0009-0004-1400-8027
ENSIAS, SSLAB, Mohammed V University, Rabat, Morocco

Amine Berqia
ENSIAS, SSLAB, Mohammed V University, Rabat, Morocco

Amol D. Vibhute
Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International University, Pune, India

Yassine Mouniane
https://orcid.org/0000-0002-6037-5301
Natural Resources and Sustainable Development laboratory, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Balakrishnama Manohar
https://orcid.org/0000-0002-8929-8709
Vellore Institute of Technology, India

ABSTRACT

Assessing the aqueous solubility of chemical compounds is one of the main tasks of molecular chemistry, with important implications for drug discovery and other research. In this respect, this study proposes a predictive model for Log S and offers a comprehensive comparative analysis of several machine learning algorithms on six different datasets, evaluating performance in terms of R-squared value, mean square error and execution time on various datasets that differ in composition and source and are used in natural and synthetic contexts. In addition, a performance comparison of the algorithms has been made in order to identify the most efficient one for solubility prediction using molecular data. Results are clearly articulated and presented, with appropriate reasoning and evidence provided. The research question

DOI: 10.4018/979-8-3693-8532-6.ch018

is appropriately addressed in the study by clarifying the differences in the response of the algorithms across six datasets of different compositions and sources. It is clear that the results can facilitate the process of comparing different machine learning algorithms.

INTRODUCTION

Data-Driven decision-making is attributed to virtue of scientific discovery and multitudes of exceptional achievements across almost all fields are made possible due to the advancement and advent of machine learning algorithms as beneficial tools **(Chaudhari et al., 2020;Sarlis et al., 2024)**. However, picking out an algorithm that is certain to be the most suitable for a certain task poses as an exceptionally daunting problem. This fact remains true because of the variety of different datasets and research objectives, as well as because of the sheer number of algorithms and their numerous distinct abilities and deficiencies **(Nisha et al., 2016;Gupta et al., 2024)**. This project serves as a comparative machine learning algorithm study, aimed at examining and comparing differed algorithms' performances against six distinct datasets. The datasets differ in size, ranging from one hundred thousand inputs to the smallest dataset of only about a hundred, along with their nature. The nature of the datasets is also bound to impact the suitable choice of algorithms in question: the molecules, proteins, and other substances possess different natures, requiring different research methods **(Po & Senozan, 2001)**. The comparison analysis was made possible through the utilization of the LazyPredict library, having automated the process of picking a model and testing its performance. The LazyPredict library outputs multiple performance evaluation metrics such as R-squared and RMSE, with the execution time included **(Bhal, 2007;Obianyo et al., 2024)**. A variety of different algorithms' performances across the datasets are studied, and the purpose of the study is to provide additional information on the performance of algorithms, which is highly useful for researchers. Such information would help interested persons to automatically determine the performance of a researched algorithm and subsequently select the most suitable one. Moreover, the purpose of the research is not singular: the act of assisting other researchers in their process of making informed decisions would allow their fields – drug discovery and molecular research – to be improved.

BACKGROUND

1. UNDERSTANDING LOG S IN MOLECULAR CHEMISTRY

The logarithm of solubility, or concept of Log S critically affects molecular chemistry. Solubility or the ability of a defined substance to dissolve in a definite solvent, is an elementary character critical to many processes involved in pharmaceutical research and drug development **(Rezaei et al., 2021)**. Log S is a specie-free compound derived from solubility that serves as a standard measure facilitating the necessary comparison of different compounds. Its main value is that Log S facilitates the prediction of the functioning of a defined molecule in the organism, namely, its ability to affect the absorption, distribution, metabolism, and excretion processes. As the process has a significant impact on the functionality and operational characteristics of different compounds, rational drug design and the formation