WAB

Provadis School of International Management and Technology

Exposé

Examining Random Forest vs Neural Networks for LogP Prediction of Drug-like Compounds

A Proof-of-Concept Implementation and Performance Analysis

Rubin Chempananickal James
rubin.chempananickal-james@stud-provadis-hochschule.de
Matriculation Number: D876

Department: Information Technology
Module: Fortgeschrittene Programmierung
Reviewer: Prof. Dr. Henrik Paul

February 18, 2026

Contents

Glossary

logP Logarithm of the octanol-water partition coefficient. A measure of a compound's hydrophobicity, or how well it dissolves in fats versus water.. 1

ZINC20 ZINC Is Not Commercial (2020), a free database of commercially available compounds for virtual screening. 1

Exposé

Problem Statement

The accurate prediction of logP is crucial in various fields, including drug discovery and environmental chemistry.

Objectives

The primary objectives of this project are:

- To find a curated dataset of molecules and find whether a fingerprint-based logp estimation is feasible
- To evaluate the performance of the implemented models in terms of accuracy and efficiency.

Methodology

The project will follow these steps:

1. Data Collection: Gather a dataset of chemical compounds with known logP values. The ZINC20 database will be used as a primary source for this data.
2. Fingerprint Generation: The molecules will be converted to fingerprints using multiple algorithms, like Morgan (ECFP4) and MACCS Fingerprints.
3. Model Development: Train multiple machine learning models like Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN)

Planned Structure

The paper will likely be structured as follows:

- Introduction
- Research Question and Objectives

- Literature Review
- Methodology
- Results and Discussion
- Limitations
- Conclusion and Future Work
- References
- AI Declaration
- Declaration of Authorship

Bibliography

Baikété, Juda, Alhadji Malloum, and Jeanet Conradie (2026). „Comparative study of machine learning methods for accurate prediction of logP and pKb“. In: *Artificial Intelligence Chemistry* 4.1, p. 100101. ISSN: 2949-7477. DOI: https://doi.org/10.1016/j.aichem.2025.100101. URL: https://www.sciencedirect.com/science/article/pii/S2949747725000181.

Irwin, John J. and Brian K. Shoichet (2005). „ZINC - A Free Database of Commercially Available Compounds for Virtual Screening“. In: *Journal of Chemical Information and Modeling* 45.1. PMID: 15667143, pp. 177–182. DOI: 10.1021/ci049714+. eprint: https://doi.org/10.1021/ci049714+. URL: https://doi.org/10.1021/ci049714+.