Integrating Morgan Fingerprints and Physicochemical Descriptors for Solubility Prediction

Vasudha Pai Karkala[1], Mirna Ashri[1], Patrick Mathieu Laborbe[1]
[1]NOVA School of Science and Technology, NOVA-FCT, 2829-516 Caparica, Portugal
v.karkala@campus.fct.unl.pt; m.ashri@campus.fct.unl.pt; l.mathieu@campus.fct.unl.pt

Aqueous solubility is a critical parameter in drug discovery, yet reliable prediction remains challenging. We developed hybrid machine-learning models combining Morgan fingerprints with six physicochemical descriptors (MolLogP, TPSA, Fsp^3, aromatic ring count, HBD, and HBA) to predict aqueous solubility on a merged ESOL + AqSolDB dataset comprising approximately 9,980 compounds. Random Forest and XGBoost models were trained and evaluated using an 80/20 train-test split. The hybrid representation consistently outperformed fingerprint-only and descriptor-only models, achieving R^2 = 0.783 and RMSE = 1.116 logS units. Feature-importance analyses identified MolLogP, TPSA, and Fsp^3 as key contributors to model predictions, in agreement with established physicochemical determinants of solubility reported in the literature. The final model was deployed as a Streamlit web application for rapid solubility prediction from user-provided SMILES strings.