Machine Learning-Driven Retention Time Prediction for High-Throughput Evaluation of Supercritical Chromatography Columns

Sakthivel Balasubramaniyan
Hokkaido University

Chromatography is essential for accurately separating and analyzing complex mixtures in environmental studies, food sciences, and pharmaceuticals. While column selection greatly influences chromatographic efficiency, extensive experimental testing of new columns can be resource-intensive due to unpredictable retention behaviors. This study addresses these challenges by employing supercritical fluid chromatography (SFC) coupled with machine learning (ML) algorithms implemented in the DOPtools library to predict retention times (RT) from molecular structures, significantly improving column evaluation efficiency. An automated robotic system synthesized a diverse library of 64 amides, whose retention times were measured using a DCpak® PBT column. These chemical structures were encoded using binary fingerprints and fragment descriptors. Regression models utilizing Support Vector Machines (SVM), Random Forests (RF), and XGBoost (XGB) were benchmarked through repeated five-fold cross-validation. Fragment descriptors combined with SVM notably outperformed binary fingerprints, particularly for compounds with high RT. The optimal model precisely predicted logarithmic retention factors (lnRF). External validation on chemically diverse test sets confirmed the model's applicability domain and identified outliers linked to novel substructures, which were also interpreted via ColorAtom analysis. This ML-driven methodology accelerates column screening, reduces experimental trial and error, and enhances analytical throughput by guiding the strategic selection of new SFC media.