Integrating LSER Descriptors and Machine Learning for IAM Chromatography

Wiktor Nisterenko[1], Katarzyna Ewa Greber[1], Magdalena Kierkowicz[2], Krzesimir Ciura[1]

[1] *Department of Physical Chemistry, Faculty of Pharmacy, Medical University of Gdansk, Al. Gen. J. Hallera 107, 80-416, Gdansk, Poland*

[2] *UCB, Slough, UK*

Understanding how molecular structure affects chromatographic retention is essential for optimizing separation conditions and predicting compound behavior. The Linear Solvation Energy Relationship (LSER) framework has been a cornerstone in this field, describing retention through physically meaningful molecular descriptors that represent specific intramolecular interactions.[1] In this study, we demonstrate how combining LSER with modern machine learning methods can enhance both predictive accuracy and interpretability.

We used immobilized artificial membrane (IAM) chromatography as our model system, which mimics biological membranes and is widely used in drug discovery to assess membrane permeability.[2,3] Our dataset consisted of 993 structurally diverse compounds with experimentally measured retention values (CHI_{IAM}).[4,5] LSER descriptors were calculated using Absolv software and extended to include ionization state, making this approach applicable to realistic drug-like molecules that exist in different ionization forms at physiological pH.

We systematically evaluated several regression algorithms, including linear models, k-nearest neighbors, support vector regression (SVR), and ensemble methods like random forest and gradient boosting. Support vector regression with a radial basis function kernel (SVR-RBF) showed the best overall performance, achieving R^2 values of 0.884 (training), 0.853 (test), and 0.811 (cross-validation). The prediction errors (RMSE between 4.6 and 7.0) were close to experimental uncertainty, and the consistent performance across validation sets indicates good generalization without overfitting.

To address interpretability, we applied SHAP analysis, which quantified each descriptor's contribution to the predictions. The analysis confirmed that hydrophobic volume (V descriptor) and hydrogen-bonding properties (A and B descriptors) are the main factors controlling IAM retention – exactly what we expected from membrane partitioning theory.

We also defined the model's applicability domain using a Williams plot (±3 standardized residuals, leverage threshold h*), establishing clear boundaries for reliable predictions. This is crucial for practical applications, as it prevents overconfident predictions for compounds outside the training space.

Our results show that integrating physically meaningful LSER descriptors with machine learning creates a powerful tool that is both accurate and interpretable. This approach can support virtual screening efforts, guide property-based molecular design, and reduce experimental workload in drug discovery. By maintaining mechanistic grounding while leveraging modern algorithms, we bridge traditional understanding with computer power, offering a practical solution for modeling membrane-mimetic chromatographic systems.

Bibliography :

[1] M.H. Abraham; A. Ibrahim; A.M. Zissimos. J. Chromatogr. A 1037 (2004) 29–47.

[2] K. Valko; M. Plass; C. Bevan; D. Reynolds; M.H. Abraham. J. Chromatogr. A 797 (1998) 41–55.

[3] K. Valko; C.M. Du; C.D. Bevan; D.P. Reynolds; M.H. Abraham. J. Pharm. Sci. 89 (2000) 1085–1096.

[4] K. Ciura. J. Chromatogr. A 1714 (2024) 464549.

[5] W. Nisterenko; B. Żołnowska; J. Deng; D. Zgoda; et al. J. Chromatogr. A 1751 (2025) 465911.