Bayesian Multilevel Modeling of Retention Data Informed by Structural Similarity of Analytes

Paweł Wiczling\*

Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Al. Gen. Hallera 107, 80-416 Gdańsk, Poland

ABSTRACT: When modeling chromatographic retention time data for heterogeneous analytes, a common assumption is that the retention behavior of individual analytes is independent (controlling for the effect of various descriptors). However, in practice, analytes often share structural similarities that introduce dependencies, which can improve predictive accuracy when leveraged. In this study, I present a multilevel modeling approach that incorporates a Tanimoto similarity matrix to address these relationships. I evaluate the proposed model using a publicly available dataset of isocratic reversed-phase high-performance liquid chromatography retention time measurements for 1,026 analytes. Many of these analytes exhibit high structural similarity, making them ideal for investigating the advantages of incorporating similarity into models. A central component of the model is its use of a matrix normal distribution parameterized by a mean matrix and two covariance matrices, one capturing covariance across analytes and the other capturing covariance across chromatographic parameters. The mean matrix includes molecular predictors, such as *logP* values and the number of functional groups. The row covariance matrix is structured according to the Tanimoto similarity matrix, which enables structural similarities among analytes to inform retention predictions. The column covariance matrix captures dependencies among model parameters. This study demonstrates that structural similarity can be seamlessly integrated into the retention time model to offer improved predictive performance, especially when experimental data from structurally related analytes is available.

Intuitively, it is clear that knowing the retention time of imipramine, a tricyclic antidepressant, can help predict the retention time of desipramine, another structurally similar tricyclic compound. However, imipramine's retention time offers little predictive value for ribitol, a sugar alcohol with no structural resemblance. One strategy to improve prediction accuracy in this setting is to use the localized QSRR approach1,2, which focuses only on structurally similar compounds while excluding dissimilar ones, such as ribitol. However, this method sacrifices generalizability because it underutilizes the full dataset, as even dissimilar compounds can inform model parameters. From a multilevel modeling perspective3–9, each analyte retention is predicted based on population-level parameters shared across all analytes and individual-level observations. Ideally, a retention time prediction model would incorporate all available analytes and weight each according to structural similarity. This hierarchical framework would enable effective information sharing across compounds and improve predictive accuracy by balancing individual- and population-level information. A multiple output Gaussian process model using a matrix normal distribution is a convenient way to extend standard multilevel models to handle analyte similarity.

Building on previous efforts to model a publicly available dataset3,4, this study focuses specifically on assessing the benefit of incorporating Tanimoto structural similarity as a predictor of covariance between analyte pairs. The data were analyzed using a multilevel approach with a normal matrix distribution. This distribution was parameterized by a mean matrix and two covariance matrices. The first matrix captured covariation across analytes, and the second matrix captured covariation across chromatographic parameters. The mean matrix incorporates predictors such as log P values and the number of functional groups. The column covariance matrix modeled the dependency between *logk­w* and *S1* of the Neue equation, and the row covariance matrix was correlated with the Tanimoto similarity matrix. This allowed structural similarities between analytes to inform model based predictions. This methodology's benefit is evaluated by quantifying how much uncertainty the similarity matrix explains and how this enhances predictive accuracy when retention data for structurally similar analytes are available.

# EXPERIMENTAL SECTION

## Data

In this work I used a publicly available datast10 that comprises the measurements of reversed-phase high-performance liquid chromatography (RP-HPLC) retention times collected for 1026 analytes. The retention times were measured under isocratic conditions on Eclipse Plus C18 (Agilent) stationary phase with 3.5 μm particles. The experiments were conducted using a mixture of two solvents: solvent A, which was made of 0.1% formic acid in water, and solvent B, which was made of 0.1% formic acid in acetonitrile. The column temperature was set at 35°C. The data were collected by Boswell et al. and were used to create a method to predict retention time by Back-Calculating the Gradient.11,12

The two-dimensional (2D) structures of the analytes, the Tanimoto similarity matrix, the log P and the number of functional groups (*X*) were calculated using the RDkit toolkit13: based on the analytes structures generated from SMILES strings. The *X* was then divided into non-dissociated (*X1*) and dissociated (*X2*) functional groups at the experimental pH.

The raw data used for the selected analytes are shown in Figure S1.

## Structural Model

## The logarithm of retention factor (*logki,j*) was modeled using the Neue model:14

## where *j* denotes observation, *i* denotes analyte, represents the logarithm of retention factors extrapolated to 0% of organic modifier content, and are the slopes in the Neue equation. In this parametrization of the Neue equation, reflects the difference between logarithms of retention factors corresponding to water (0% organic modifier) and 100% organic modifier content as eluents. For simplicity the *S2* parameter was assumed identical for all the analytes (*S2,i* = .

## The statistical model has the following hierarchical structure:

## where *Ri* = () is a vector of analyte-specific parameters, corresponds to the Neue equation above, MN denotes a matrix normal distribution, *θR* is a vector of typical values of *Ri* (*θlogkw, θS1*) for an analyte with *logP* equal 2.2 and no functional groups, *θdR* is a vector denoting the effect of dissociating functional group on *Ri*, *βR* is a vector of slopes with respect to *logPi*, π1R and π2R are vectors of slopes for the non-dissociated and dissociated functional groups. Parameters *νobs*=7 and σ denote the normality and scale of residuals, respectively.

## *K* and are the scale matrices for the row and column covariance structures of the matrix normal distribution, decomposed as:

## where *LL'* is a correlation matrix, and *L* is a Cholesky of the correlation matrix, *ω* is a vector of standard deviations for between analyte variability of *logkw* and *S1* values, *S* is a Tanimoto similarity matrix, and *I* is the identity matrix. The similarity matrix was simplified by setting *S* = 0 for S < 0.5, treating those analytes as uncorrelated, and scaling values *S* ≥ 0.5 by α using the above formula. Value of 0.5 was selected arbitrary as a similarity cut-off.

## MN distribution models random matrices using a mean matrix and two covariance matrices: *i*) one for rows (analytes) and *ii*) one for columns (for *logkw* and *S1*). These covariance matrices *K* (row) and *Ω* (column) describe dependencies across analytes and between parameters, respectively. The row covariance *K* is correlated with analyte similarity. This setup ensure that the values of *logkw* and *S1* across analytes are jointly Gaussian, and the way they covary is proportional to a known similarity measure. As a result, both inter-analyte relationships and parameter correlations are effectively modeled.

Further, to simplify computation, analytes were grouped based on Tanimoto similarity using community detection (cluster\_louvain()), resulting in clusters of structurally similar compounds. This enabled the partitioning of analytes into blocks, allowing operations on smaller submatrices. The approach significantly reduced computational burden, as Cholesky decompositions of large matrices (*n* > 1000) are typically slow and resource-intensive.

## Priors

The Bayesian model requires specification of priors that in this study had the following form:

*LN* denotes lognormal distribution, *N+* denoted half normal distribution, *T*[0,1] indicates that the distribution is truncated to the interval [0, 1], and *LKJ*(2) ensure that the density is uniform over correlation matrices of order 2. These priors are weakly informative ensuring regularization of model parameters.

**Bayesian Inference**

**Technical.** Multilevel modeling was performed using the Stan/cmdstanr15 software linked with Rstudio16. For the inference, we used eight Markov chains with 500 iterations after 1000 warm up iterations. Convergence diagnostics were performed using Gelman–Rubin statistics and trace plots, and the results indicated that the model results did not diverge. The R code, data, and Stan code used to analyze the data are publicly available in the GitHub repository (https://github.com/wiczling/izocratic-qsrr). The calculations were run on the Tryton computing cluster in the Center of Informatics Tricity Academic Supercomputer and Network.

**Diagnostics.** The individual-level random effects were decorrelated to check if the model correctly captured the correlation structure using the following equation

It first removes scaling effects using the Cholesky factor of the row covariance matrix, then removes correlation using the Cholesky factor of the column covariance. The resulting decorrelated values should follow a standard normal distribution if the model is specified well.

**Predictions.** The applicability of the proposed model was demonstrated through the prediction of analyte retention conditional on the retention of structurally similar compounds. A cross-validation procedure was employed to evaluate its performance, with cross-validation folds constructed so that at least one similar analyte remained in the training set for each held-out analyte. Analytes that did not belong to any correlated group (i.e., unclustered analytes) were always included in the training set.

A similar cross-validation strategy was used to assess predictive uncertainty under the assumption that only one observation was available for each held-out analyte, reflecting a realistic scenario in which limited retention data (e.g., from a scouting experiment) is available to inform predictions.

# RESULTS AND DISCUSSION

In this study, we used an extension of the previously developed multilevel models.4–8 The key model components are (*i*) a measurement error model, (*ii*) a model for analyte-specific chromatographic parameters (*logP* and functional groups further divided into dissociated and non-dissociated groups), and (*iii*) a model for functional group effects (regression coefficients for functional group effects). The new component comprise a more detailed modeling of analyte parameters covariances and their relationship to similarity matrix.

The key population level model parameters are summarized in Table S1 and Figures 1.

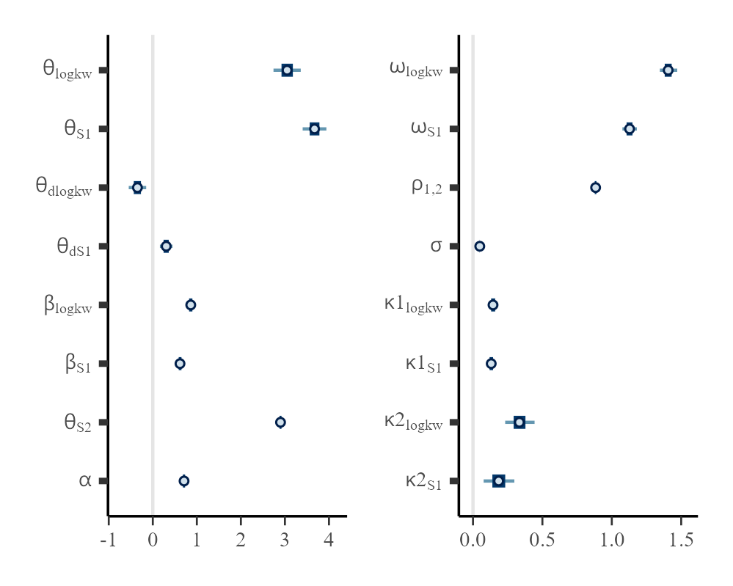


Figure 1. Summary of the marginal posterior distributions of the key population-level parameters.

Typical estimates of the *logkw* and *S1* parameters for an analyte with *logP* equal 2.2 and no functional groups are 3.05 and 3.67, respectively. The *S2* parameter is relatively high, with a value of 2.90. On average, the presence of a dissociated group decreases *logkw* by 0.35 and increases *S1* by 0.30, indicating meaningful influence of ionization on these parameters.The estimated effect of *logP* is 0.86 for *logkw* and 0.62 for *S1*, reflecting a significant contribution of hydrophobicity to both retention parameters.

The standard deviation for between-subject variability (BSV) was estimated at 1.41 for *logkw* and 1.13 for *S1*. There is a strong positive correlation of 0.88 between the two, indicating substantial shared variability across analytes. The parameter rho-squared (*ρ*²) can be interpreted as the proportion of variance in predicted *logk* explained by a single *logk* measurement, and it was estimated at approximately 0.8. The residual error (*σ*) is small at 0.05, consistent with expectations for retention time data. This corresponds to a coefficient of variation (CV) for the retention factor (*k*) of approximately 11%. Finally, the α parameter was estimated at 0.71, suggesting that, for highly similar analytes, the effective correlation is somewhat reduced. The squared term, *α²*, can be interpreted as the proportion of variance in the predicted *logk* explained by the retention of a similar analyte with a similarity score of 1. This represents a substantial level of explained uncertainty.

The effects of functional groups (population-level parameters) are shown in Figure 2 for the most common groups in the dataset, and in Figure S2 for the less common ones. The presence of dissociated functional groups, such as aliphatic amine, amidine, and guanidine, tends to significantly decrease *logkw* and increase *S1*. Non-dissociated groups generally have smaller effects, though several (e.g., annelated rings, secondary alcohols, primary carbons) still show non-negligible deviations from zero. Figure 1 reports the κ values, which summarize variability across functional groups: 0.15 and 0.13 for non-dissociated groups, and 0.34 and 0.19 for dissociated groups, for *log kw* and increase *S1*, respectively. These values suggest that functional groups not dissociated under the experimental pH conditions are expected to affect predictions by approximately ±0.30 (2 standard deviations). This component of the model supports generalization to analytes containing functional groups not represented in the analyzed dataset.

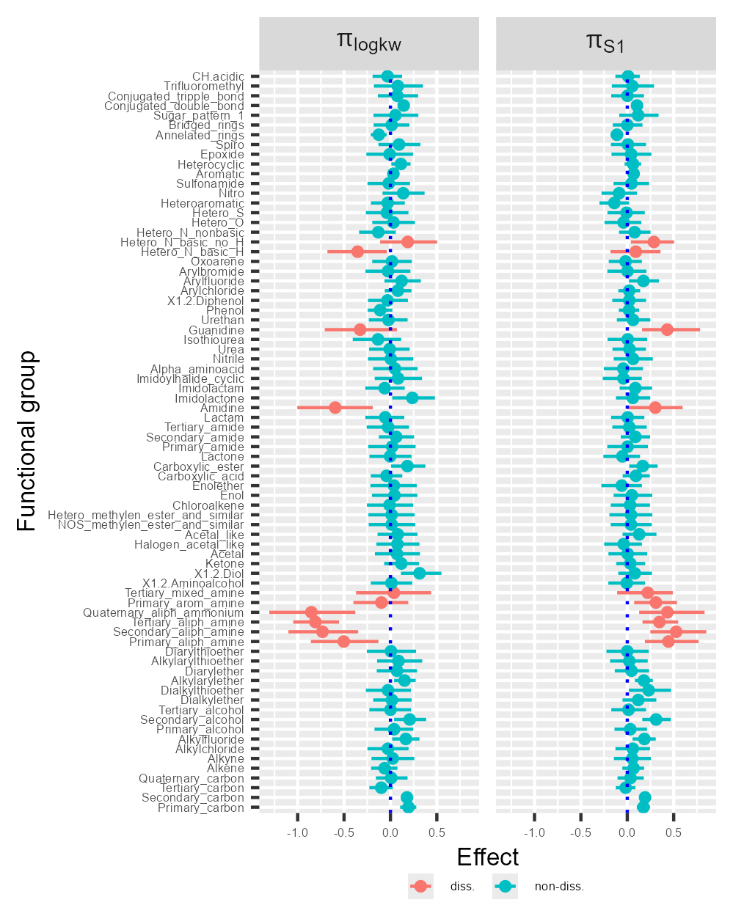


Figure 2. This is a summary of the marginal posterior distributions of the functional group effects. Red represents the effect of dissociated functional groups, and green represents the effect of non-dissociated functional groups. The most common functional groups (>10 across all analytes) are presented. The plot for the remaining functional groups is presented in Figure S2.

The distributions of analyte-specific chromatographic parameters are shown in Figures S4 and S5, either directly or as eta plots, which capture unexplained variability. Here, eta represents the deviation of an individual analyte's parameter from its expected (population-level) value. One way to assess the influence of the similarity matrix on predictions is by examining the posterior correlations between individual eta values. The presence of such correlations indicates shared information between analytes. Incorporating the similarity matrix reduces these correlations, particularly for highly similar analytes, as illustrated in Figure 3. In this figure, the high posterior correlation for similarities higher the >0.5 are significantly reduced (Figure 3A vs Figure 3B) supporting the proposed row covariance structure of the matrix normal distribution.

The model predictions are well calibrated with the observed data, as shown in Figure S3. Both individual and population-level predictions are symmetrically distributed around the line of identity, indicating good overall agreement. Individual predictions are highly precise and closely match the observed values. Population predictions are also well calibrated, though, as expected, they are less precise due to the absence of individual-level information.

Figure 4 shows cross-validation-based predictions, which demonstrate a moderate to substantial reduction in prediction uncertainty compared to population-based predictions. This reduction is summarized for two scenarios: one using population-level predictions conditional on analytes included in the held-out data and with no individual-level observations, and another using limited data predictions with a single observation of held-in analytes. The extent of variance reduction varies widely depending on the analyte's similarity to others. Figure 5A shows that analytes with similarity scores between 0.8 and 1 experienced a notable reduction in *logk* uncertainty, typically in the range of 20–60%. The presence of a single observation reduces variance by an additional 80% due to the correlation between *logkw* and *S1* (Figure 4A vs. Figure 5B).

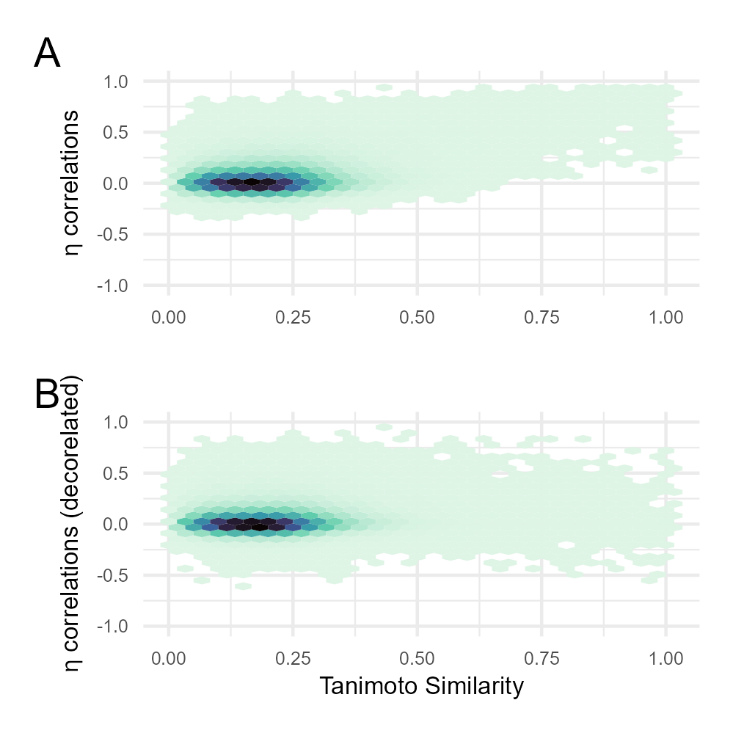


Figure 3. The hexbin heatmap presenting the relationship between eta posterior correlations and similarity matrix for every pair of analytes (A). If the model correctly handles individual-level correlations, one expects to see almost no trend in the graphs with decorrelated etas (B).

The proposed multilevel modeling approach contrasts with the approach proposed by Haddad et al.1,2, known as localized quantitative QSRR modeling. Their method identifies analytes similar to the target analyte from an available database using predefined similarity measures, such as structural similarity (e.g., Tanimoto index), physicochemical properties (e.g., lipophilicity), or acid/base character. A QSRR model is then built on this subset and used for prediction. While this approach tailors predictions to the local chemical context, it can suffer from instability when few similar compounds are available and it omits the available information about the retention present in the removed data . In contrast, the proposed multilevel model is a more generally approach that inherently balances local specificity with global information sharing. In addition the Bayesian component quantifies uncertainty that is relevant when used for predictions when limited experimental information is available.

Accurate predictions (with well-constructed and calibrated uncertainties) are critically important17, and this model further advances efforts to develop more reliable and precise prediction tools in chromatography. Since the model handles uncertainties it can also be effectively used to solve various decision problems encountered by chromatographer18–21.

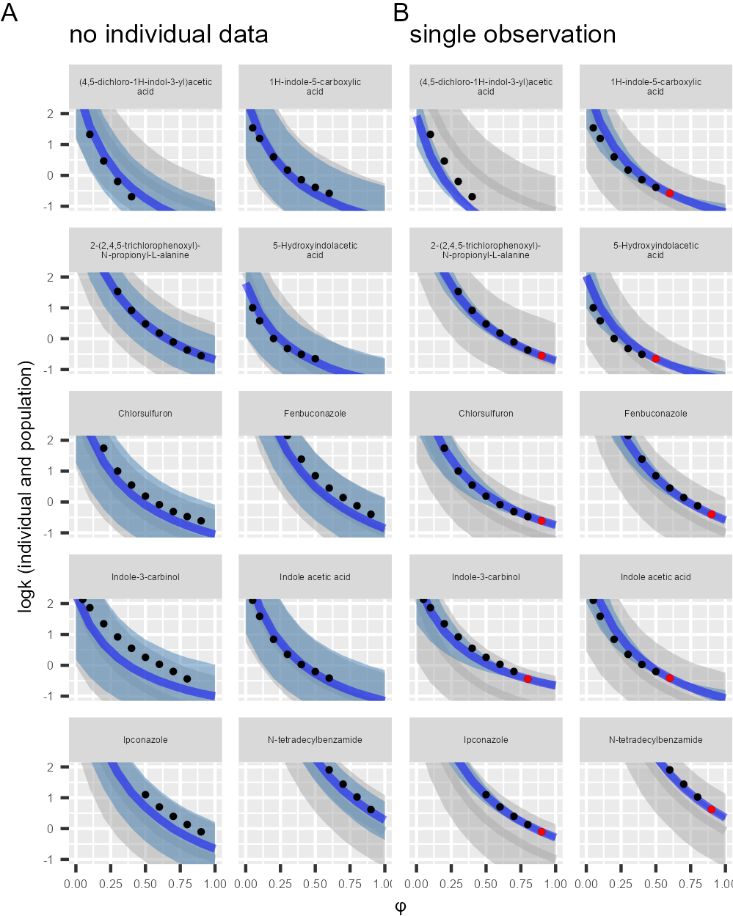


Figure 4. The predictions are represented as the posterior median (line) and the 5th–95th percentiles (shaded areas) for ten exemplary analytes. The gray area represents population predictions for future observations based solely on population-level parameters and predictors, with no similar analytes available. The blue area represents individual predictions, conditional on retaining other analytes (A) and on single observed data (the smallest observed log k, depicted as a red dot) (B). The narrower blue lines indicate the added predictive value of the similarity matrix alone (A) and the similarity matrix plus the single observation (B).

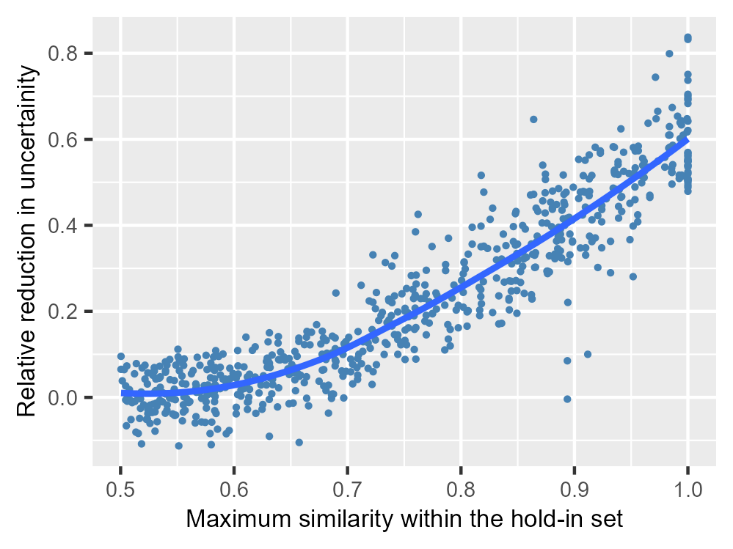


Figure 5. The relationship between the relative reduction in uncertainty and maximal similarity within the hold-in set of analytes determined using cross-validation. Only analytes that have at least one similar analyte (*S* > 0.5) are shown.

# CONCLUSIONS

# This work focuses on the multilevel modeling of retention data, incorporating structural similarity between analytes. The usefulness of the proposed approach is demonstrated using a publicly available dataset of isocratic RP-HPLC retention times for 1,026 analytes, several of which exhibit high structural similarity. The work shows that structural similarity can be easily calculated from SMILES representations and can help predict analyte retention when experimental data on structurally similar analytes is available. A reduction of up to 60% in the uncertainty of the logarithm of the retention factor is expected when retention data for highly similar analytes is available.

# ASSOCIATED CONTENT

## Supporting Information

The following Supporting Information is available free of charge at the ACS website:

# 1.  Table S1. Summary of the MCMC simulations of the marginal posterior distributions of population-level model parameters; Figure S1. Raw data; Figure S2. Functional group effects; Figure S3. GOF; Figure S4. Individual parameters; Figure S5. Eta plots; Figure S6. Compare limited data predictions.

# AUTHOR INFORMATION

## Corresponding Author

\* wiczling@gumed.edu.pl

## Author Contributions

PW conceived the presented idea, analyzed the data and wrote the paper.

**Notes**  
The author declare no competing financial interest.

# FUNDING

# This work was supported entirely by departmental funds

# ACKNOWLEDGMENT

To my wife, for her constant support.

# REFERENCES

(1) Wen, Y.; Talebi, M.; Amos, R. I. J.; Szucs, R.; Dolan, J. W.; Pohl, C. A.; Haddad, P. R. Retention Prediction in Reversed Phase High Performance Liquid Chromatography Using Quantitative Structure-Retention Relationships Applied to the Hydrophobic Subtraction Model. *Journal of Chromatography A* **2018**, *1541*, 1–11. https://doi.org/10.1016/j.chroma.2018.01.053.

(2) Park, S. H.; Talebi, M.; Amos, R. I. J.; Tyteca, E.; Haddad, P. R.; Szucs, R.; Pohl, C. A.; Dolan, J. W. Towards a Chromatographic Similarity Index to Establish Localised Quantitative Structure-Retention Relationships for Retention Prediction. II Use of Tanimoto Similarity Index in Ion Chromatography. *Journal of Chromatography A* **2017**, *1523*, 173–182. https://doi.org/10.1016/j.chroma.2017.02.054.

(3) Wiczling, P.; Kamedulska, A.; Kubik, Ł. Application of Bayesian Multilevel Modeling in the Quantitative Structure-Retention Relationship Studies of Heterogeneous Compounds. *Anal Chem* **2021**, *93* (18), 6961–6971. https://doi.org/10.1021/acs.analchem.0c05227.

(4) Kamedulska, A.; Kubik, Ł.; Wiczling, P. Statistical Analysis of Isocratic Chromatographic Data Using Bayesian Modeling. *Analytical and Bioanalytical Chemistry* **2022**, *414* (11), 3471–348. https://doi.org/10.1007/s00216-022-03968-x.

(5) Kamedulska, A.; Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M. J.; Wiczling, P. Toward the General Mechanistic Model of Liquid Chromatographic Retention. *Anal. Chem.* **2022**, *94* (31), 11070–11080. https://doi.org/10.1021/acs.analchem.2c02034.

(6) Wiczling, P. Analyzing Chromatographic Data Using Multilevel Modeling. *Anal Bioanal Chem* **2018**, *410* (16), 3905–3915. https://doi.org/10.1007/s00216-018-1061-3.

(7) Wiczling, P.; Kamedulska, A. Comparison of Chromatographic Stationary Phases Using a Bayesian-Based Multilevel Model. *Anal Chem* **2024**, *96* (3), 1310–1319. https://doi.org/10.1021/acs.analchem.3c04697.

(8) Kubik, Ł.; Kaliszan, R.; Wiczling, P. Analysis of Isocratic-Chromatographic-Retention Data Using Bayesian Multilevel Modeling. *Anal Chem* **2018**, *90* (22), 13670–13679. https://doi.org/10.1021/acs.analchem.8b04033.

(9) Wiczling, P.; Kubik, Ł.; Kaliszan, R. Maximum A Posteriori Bayesian Estimation of Chromatographic Parameters by Limited Number of Experiments. *Anal Chem* **2015**, *87* (14), 7241–7249. https://doi.org/10.1021/acs.analchem.5b01195.

(10) *retentionprediction.org/hplc - accurate HPLC retention prediction*. https://www.retentionprediction.org/hplc/database/ (accessed 2025-07-17).

(11) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. A Study on Retention “Projection” as a Supplementary Means for Compound Identification by Liquid Chromatography–Mass Spectrometry Capable of Predicting Retention with Different Gradients, Flow Rates, and Instruments. *Journal of Chromatography A* **2011**, *1218* (38), 6732–6741. https://doi.org/10.1016/j.chroma.2011.07.105.

(12) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. Easy and Accurate High-Performance Liquid Chromatography Retention Prediction with Different Gradients, Flow Rates, and Instruments by Back-Calculation of Gradient and Flow Rate Profiles. *Journal of Chromatography A* **2011**, *1218* (38), 6742–6749. https://doi.org/10.1016/j.chroma.2011.07.070.

(13) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; Gedeck, P.; Jones, G.; Kawashima, E.; NadineSchneider; Nealschneider, D.; Dalke, A.; Swain, M.; Cole, B.; tadhurst-cdd; Turk, S.; Savelev, A.; Vaucher, A.; Wójcikowski, M.; Take, I.; Walker, R.; Scalfani, V. F.; Faara, H.; Ujihara, K.; Probst, D.; Maeder, N.; Monat, J.; Lehtivarjo, J.; godin, guillaume. Rdkit/Rdkit: 2025\_03\_4 (Q1 2025) Release, 2025. https://doi.org/10.5281/zenodo.15773589.

(14) Nikitas, P.; Pappa-Louisi, A. Retention Models for Isocratic and Gradient Elution in Reversed-Phase Liquid Chromatography. *Journal of Chromatography A* **2009**, *1216* (10), 1737–1755. https://doi.org/10.1016/j.chroma.2008.09.051.

(15) Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. Stan: A Probabilistic Programming Language. *Journal of statistical software* **2017**, *76* (1).

(16) RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio, PBC.: Boston, MA, 2020.

(17) Gritti, F. Perspective on the Future Approaches to Predict Retention in Liquid Chromatography. *Anal. Chem.* **2021**, *93* (14), 5653–5664. https://doi.org/10.1021/acs.analchem.0c05078.

(18) Wiczling, P. Evaluation of Sequential Bayesian-Based Method Development Procedures for Chromatographic Problems Involving One, Two, and Three Analytes. *SEPARATION SCIENCE PLUS*, 2018, *1*, 63–75. https://doi.org/10.1002/sscp.201700037.

(19) Yamamoto, Y.; Yajima, T.; Kawajiri, Y. Uncertainty Quantification for Chromatography Model Parameters by Bayesian Inference Using Sequential Monte Carlo Method. *Chemical Engineering Research and Design* **2021**, *175*, 223–237. https://doi.org/10.1016/j.cherd.2021.09.003.

(20) Bosten, E.; Pardon, M.; Chen, K.; Koppen, V.; Van Herck, G.; Hellings, M.; Cabooter, D. Assisted Active Learning for Model-Based Method Development in Liquid Chromatography. *Anal. Chem.* **2024**, *96* (33), 13699–13709. https://doi.org/10.1021/acs.analchem.4c02700.

(21) Wiczling, P. Leveraging Prior Knowledge for Improved Retention Prediction in Reversed-Phase HPLC. *J Chromatogr A* **2025**, *1746*, 465787. https://doi.org/10.1016/j.chroma.2025.465787.

TABLE OF CONTENTS (TOC)

