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: A common assumption when modeling chromatographic retention time data of heterogeneous analytes is that the between analyte variability across analytes is unrelated. While this assumption holds for dissimilar analytes, in practice, analytes often share structural or chemical similarities that induce dependency that can provide valuable information for improving predictive accuracy. In this work, I propose a model that incorporates the Tanimoto similarity matrix to capture these relationships and demonstrate how this approach enhances the precision of retention time predictions. The proposed approach was demonstrated using a publicly available dataset comprising isocratic RP-HPLC retention time measurements for 1,026 analytes, among which several analyte pairs exhibit high structural similarity. A key component of the model is the use of a matrix normal distribution, parameterized by a mean matrix and two covariance matrices: one capturing variation across analytes and the other across chromatographic parameters. The mean matrix incorporates predictors such as log P values and number of functional groups. Notably, the row covariance matrix is aligned with the Tanimoto similarity matrix, allowing structural similarities between analytes to inform the model and enhance predictive performance. This works demonstrates that structural similarity can be easily incorporated into the modeling workflow helping predicting analyte retention if experimental data of structurally similar analytes is available.

Analysts often approach the prediction of retention for congeneric compounds differently than for structurally unrelated analytes. Even after accounting for functional group effects, prediction uncertainties tend to be substantially lower for congeneric compounds. This challenges a common assumption in chromatographic retention modeling that variability between analytes is independent. While this assumption may hold for chemically dissimilar analytes, in practice, many analytes share structural or chemical similarities that induce dependencies. When such relationships exist, they can be leveraged to improve predictive accuracy-particularly when retention data for similar analytes is available.

This study builds on previous efforts to model publicly available datasets.1,2, with a specific focus on assessing the benefit of incorporating Tanimoto structural similarity as a predictor to reduce variability of structurally similar analytes. The data was analysed using multilevel apporach1, and using a matrix normal distribution, parameterized by a mean matrix and two covariance matrices: *i)* one capturing variation across analytes and *ii)* the other across chromatographic parameters. The mean matrix incorporates predictors such as *log P* values and number of functional groups. The column covariance model the dependency between *log kw* and *S1* of the Neue equation and the row covariance matrix was correlated with the Tanimoto similarity matrix, allowing structural similarities between analytes to inform the model and enhance predictive performance.

# EXPERIMENTAL SECTION

## Data

We used a publicly available3 that comprises the measurements of 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.4,5

The 2D structures of analytes, Tanimoto similarity matrix, log P and the number of functional groups were calculated using RDkit toolkit6: based on the structures of the analytes generated from SMILES strings.

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

## Structural Model

## The logarithm of retention factor (*log ki,j*) was modeled using the Neue model:7

## 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.

## 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*), *θdR* is a vector of typical values for dissociation effects (*θdogkw, θdS1*), *βR* is a vector of slopes with respect to *logPi*, π1R and π2R are vectors of slopes related to non-dissociated (*X1*) and dissociated (*X2*) functional groups at the pH experiments were conducted. S2 parameter was assumed identical for all the analytes (*S2* = .

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

## where *LL'* is a correlation matrix, *ω* is a vector of standard deviations across analytes, *S* is a 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.

## Parameters *νobs*=7 and σ denote the normality and scale of residuals, respectively.

## MN distribution models random matrices using a mean matrix and two covariance matrices: i) one for rows (analytes) and ii) one for columns (for *log kw* 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.

## Priors

The Bayesian model requires specification of priors that allow incorporation of domain expertise into inferences. In this study, a fairly uninformative priors were selected.

LN denotes lognormal distribution, *N+* denoted half normal distribution, and LKJ(2) ensure that the density is uniform over correlation matrices of order 2 and u denotes the unique lower triangular elements of correlation matrix

## Bayesian Inference

**Technical.** Multilevel modeling was performed using the Stan/cmdstanr8 software linked with Rstudio9. 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 by predicting the retention of an analyte conditional on the retention of structurally similar compounds. To evaluate its performance, a standard cross-validation procedure was employed. Analytes were first grouped based on correlation using community detection (cluster\_louvain()$membership), resulting in clusters of correlated compounds. Cross-validation folds were constructed so that, for each held-out analyte, at least one similar (i.e., correlated) analyte remained in the training set. Analytes that did not belong to any correlated group (i.e., unclustered analytes) were always included in the training (held-in) set.

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

# RESULTS AND DISCUSSION

In this study, we used a simple multilevel model that aligns with our previous efforts to model chromatographic data using this approach2,10–13. The key model components are (*i*) a measurement error model, (*ii*) a model for analyte-specific chromatographic parameters (*log P* and functional groups for dissociated and non-dissociated), (*iii*) a model for functional group effects (regression coefficients for functional group effects), and (*iv*) a model between similarity matrix and row covariance of matrix normal distribution. The model was further supplement with prior information.

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

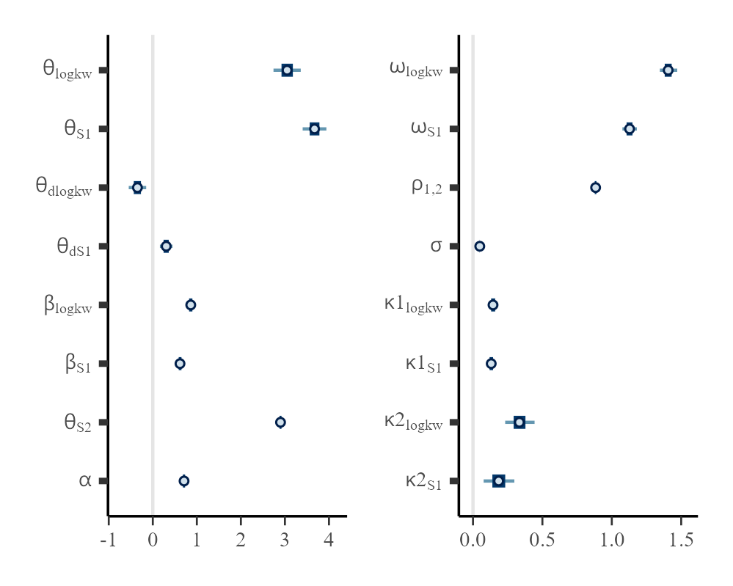


Figure 1. Summary of the marginal posterior distributions of the key population-level parameters.

The typical estimates of the *log kw* and *S1* parameters for an analyte with *log P* = 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. The presence of a dissociated group decreases *log kw* on average by 0.35 and increases *S1* by 0.30, indicating a meaningful influence of ionization on these parameters.

The effect of *log P* is estimated at 0.86 for *log kw* and 0.62 for *S1*, reflecting a strong contribution of hydrophobicity to both retention and polarity-related descriptors.

Between-subject variability (BSV) was estimated at 1.41 for *log kw* and 1.13 for *S₁*, with a strong positive correlation of 0.88 between them, indicating substantial shared variability across analytes. The parameter *ρ²* (approximately 0.8) can be interpreted as the proportion of uncertainty variance in predicted *log k* explained by a single log k measurement. The residual error (σ) is small, at 0.05, which is 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 to be 0.71, indicating that for analytes with high similarity, the effective correlation is somewhat attenuated. The squared term, *α²* (approximately 0.5), can be approximately interpreted as the proportion of variance in the predicted log k that is explained when the retention of a similar analyte (with similarity score of 1) is known. This represents a substantial level of explained uncertainty.

The functional group effect are shown on Figure 2 (for common) and Figures *S2* (for rare functional group in the dataset). The presence of dissociated function group such as aliphatic amine, amidine, guanidine decrease the *log kw* and increase *S1* considerably. The effect of non-dissociated group is smaller, but several group show effect that are different from zero (annelated ring, secondary alcohol, primary carbon).

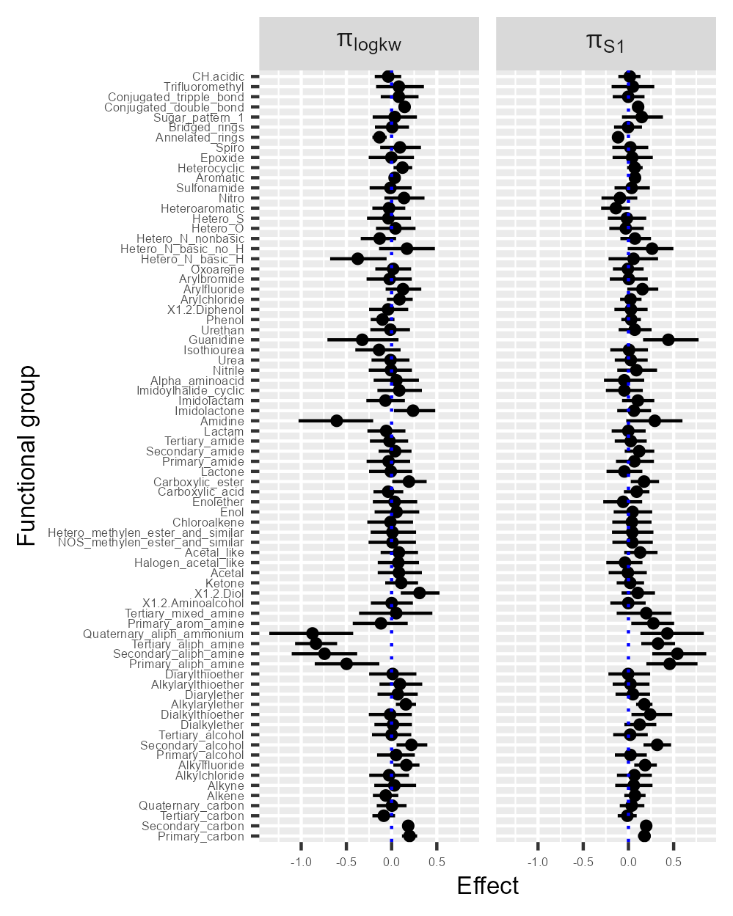


Figure 2. Summary of the marginal posterior distributions of the functional group effects. The subset of most common function groups is presented (>10 across all analytes). The plot for remaining functional groups is presented in Figure S2.

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. Explicitly modeling this similarity reduces these correlations—particularly for highly similar analytes, as illustrated in Figure 3.

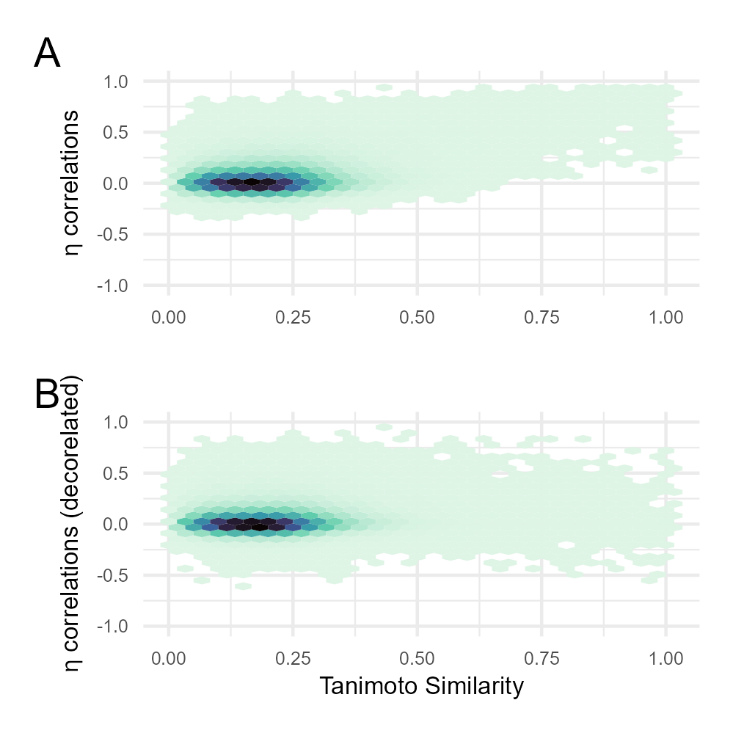


Figure 3. The hexbin heatmap presenting the relationship between eta posterior correlations and similarity matrix for every pair of analytes (A). If the model correctly handles individual-level correlations, one expects to see almost no trend in the graphs with decorrelated etas (B).

The 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.

Cross-validation-based predictions are shown in Figure 4, demonstrating 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 (i.e., without any observations) and another using limited data, specifically a single observation. The extent of variance reduction varied widely, depending on the analyte’s similarity to others. As shown in Figure 5, analytes with similarity scores between 0.8 and 1 experienced a notable reduction in retention factor uncertainty—typically in the range of 20–60%. The presence of single observation reduces the variance by additional 80% due to the correlation of *log kw* and *S1* (Figure 4A vs. Figure 5B)

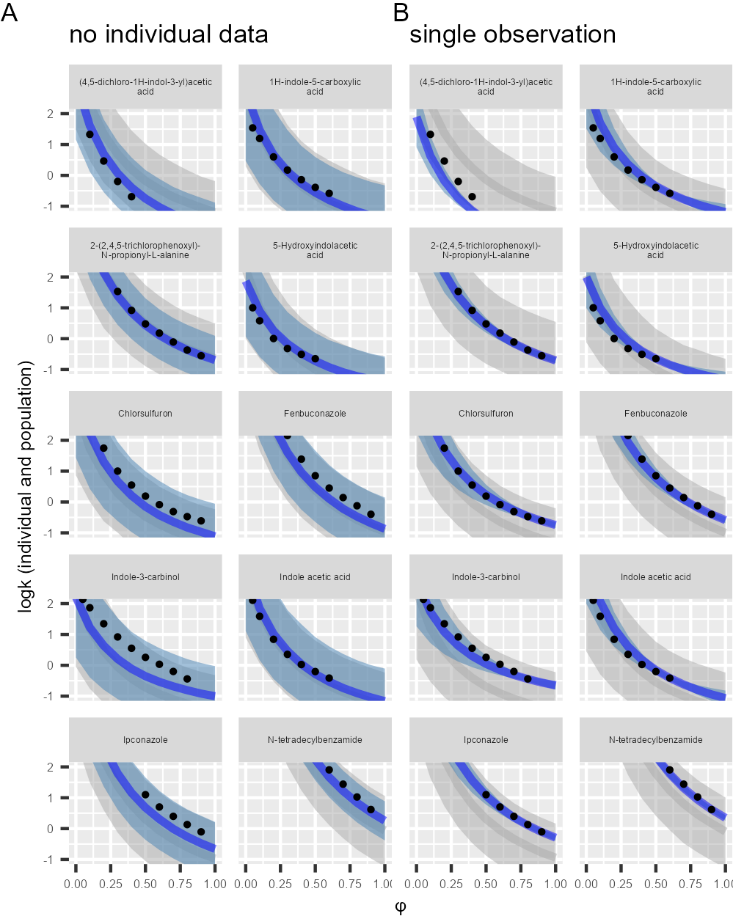


Figure 4. Predictions represented as posterior median (line) and 5th-95th percentiles (areas) for 10 exemplary analytes. The gray area represents population predictions corresponding to future observations given only population-level parameters and predictors. The blue area represents the individual predictions conditional on the retention of other analytes (A), and additionally on the single observed data (the smallest observed *log k*) (B). The narrower blue lines indicates the added predictive value of similarity matrix (A) and similarity matrix + single observation (B).

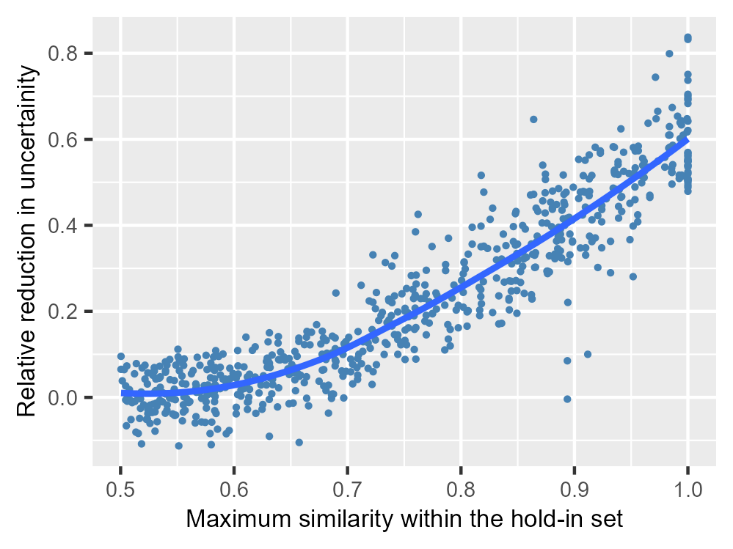


Figure 5. The relationship between the relative reduction in uncertainty and maximal similarity within the hold-in set of analytes determined using cross-validation. Only analytes that have at least one similar analyte (*S* > 0.5) are shown.

The proposed multilevel modeling approach contrasts with the approach proposed by Haddad et al.14,15, 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. 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 becomes relevant one limited information is available.

Accurate predictions are critically important16, and this model further advances efforts to develop more reliable and precise prediction tools.

# CONCLUSIONS

# This work focuses on multilevel modeling of isocratic HPLC retention data incorporating structural similarity between analytes. The usefulness of the proposed approach was demonstrated using a publicly available dataset comprising isocratic RP-HPLC retention time measurements for 1,026 analytes with several analyte pairs exhibit high structural similarity. This works demonstrates that structural similarity (such as Tanimoto similarity metric) can be easily calculated from SMILES representations and help predicting analyte retention if experimental data of structurally similar analytes is available. Up to 60% reduction of logarithm of retention factor uncertainty is expected for highly similar analytes.

# 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) 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.

(2) 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.

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

(4) 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.

(5) 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.

(6) 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.

(7) 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.

(8) 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).

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

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

TABLE OF CONTENTS (TOC)

