Comparison of chromatographic stationary phases using Bayesian-based multilevel modeling

Agnieszka Kamedulska, Łukasz Kubik, Julia Jacyna, Wiktoria Struck-Lewicka, Michał J. Markuszewski, Paweł Wiczling\*

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

ABSTRACT: We used a Bayesian multilevel model of chromatographic retention to compare five RP-HPLC stationary phases: XBridge Shield RP18, XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8. For that purpose we used a large dataset of retention times collected using chromatographic techniques coupled with mass spectrometry detection. The experiments were conducted in a gradient mode for a mixture of 300 small analytes for a wide range of pH values, in MeOH and ACN, at two temperatures, and for three gradient durations. Our analysis is based on a mechanistic model derived from principles and fundamentals of liquid chromatography and utilizes the literature prior knowledge about model parameters. The data and model allowed us to characterizes the between column differences in retention of neutral, acidic, and basic analytes. The proposed approach provides interpretable summary of stationary phase properties that can be used in decision-making, i.e. finding the best chromatographic conditions given limited experimental data.

The selection of stationary phases is an important aspect of analytical work. There are numerous methods and approaches available in literature1. The chromatographic stationary phases are usually characterized using few probe analytes at limited set of chromatographic conditions. This method provide a rough estimates of analyte characteristics on retention. Also a very detailed physical (and thermodynamic) description of the chromatographic systems are possible, nevertheless they require an extensive sets of experiments and as such are limited to few analytes2.

It also seems possible to characterize stationary phase using retention time data collected for a relatively large and heterogeneous group of compounds. Such datasets are rather easy to collect using chromatographic techniques coupled with mass spectrometry detection. There are, however, more difficult to analyze. Since such datasets carry much information about chromatographic retention they allow for a detailed multidimensional characterization of chromatographic stationary phases and ability to predict retention (along with uncertainty) based on various number of preliminary experiments (e.g. to predict retention time for a set of analytes given no, or several measurements collected using a different stationary phase). The accurate predictions of the retention time in liquid chromatography is required for rapid column screening, computer-assisted method development and method transfer3

In this work we applied the previously developed Bayesian multilevel framework4–7 to characterize chromatographic gradient retention time datasets collected using a multicomponent mixtures of analytes, five stationary phases, and a wide range of chromatographic conditions (pH, organic modifier, temperature, gradient program). The general idea is to statistically characterize the retention of acids, bases and neutral analytes using common chromatographic parameters, such as logkw and S of the Neue model.

We decided to compared five RP-HPLC stationary phases XBridge Shield RP18, XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8. At the end we illustrate the usefulness of the model for decision making given access to different types of preliminary data.

# EXPERIMENTAL SECTION

## Data

The data were collected by performing 84 different liquid chromatography experiments using an initial mixture of 300 analytes. The experiments differed with respect to gradient duration (30, 90, and 270 min), pH of the mobile phase (from 2.5 to 10.5), type of organic modifier (methanol (MeOH) or acetonitrile (ACN)) and column temperature (25 °C and 35 °C).

Briefly the liquid chromatography experiments were carried out using an Agilent Technologies 1260 Infinity system (Agilent Technologies, Waldbronn, Germany) and a 6224 time of flight (TOF) mass spectrometer with a dual electrospray ionization source (Dual ESI) in positive polarity, using an XBridge Shield RP18, XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8 Waters Ltd., Milford, MA, USA, 3 mm × 50 mm, 2.5 μm). The extra column volume and system dwell volume (*Vd*) equaled 0.020 mL and 1.05 mL, respectively. The column hold-up volume (*V0*) was 0.266 mL 0.271, 0.271, 0.276, 0.284 mL and the flow rate (*F*) was 0.5 mL/min.

Ammonium bicarbonate, ammonium acetate and ammonium formate were selected as buffers to control the pH of the mobile phase during chromatographic separation. The pH of the buffers (nominal aqueous pH) was adjusted to the desired pH (ammonium formate: 2.5, 3.3, 4.1, 8.9, and 9.7; ammonium acetate: 4.9 and 5.8; and ammonium bicarbonate: 6.8 and 10.5) by an appropriate addition of formic acid, acetic acid and ammonia, respectively. The pH was measured at 25 °C and 35 °C using an S220 pH meter (Mettler Toledo, Greifensee, Switzerland) with an InLab® Routine Pro ISM electrode after mixing an organic modifier with the buffer solution.

The MassHunter Profinder B.08.00 (Agilent Technologies, Waldbronn, Germany) was selected to find all the matches per formula using “Batch Targeted Feature Extraction” (containing 300 predefined mases for each analyte included in the mixture). The data for analysis were restricted to analytes that had “IdentificationScores” higher than 95%, that were present on at least 42 chromatograms, and that had less than 2 dissociation steps in a pH range from 2 to 11.

The functional groups and structural elements were determined using Checkmol (version 0.5b N. Haider, University of Vienna, 2003-2018).8 The lipophilicity (*log P*), dissociation constant (*pKalit*) were calculated using the ACD/Labs program9 based on the structures of analytes generated from SMILES strings.

More details about data extraction can be found in our previous work.4 Raw data for selected analytes is shown in Figure S1.

## Structural Model

The details of the model are provided in supplementary material. Briefly, a standard chromatographic model was employed in this work.10,11 For each analyte the effect of pH was accounted for by the following function describing the relationship between the isocratic retention factor and *pH* for an analyte with *R* dissociation steps and *R+1* forms12:

(1)

where *r* represents thedissociation step, denotes the *rth* dissociation constant, andrepresents the retention factor of a particular form of the analyte in a given chromatographic condition. Furthermore, it was assumed that depends on the organic modifier content, pH and temperature according to the following equation:

(2)

where represents the logarithm of retention factors extrapolated to 0% of organic modifier content at 25 °C for mobile phase pH 7 for the neutral and dissociated forms of the analyte; and denote slopes in the Neue equation; denotes the change in due to the increase in temperature by 10 °C, *apHr* denotes pH effects for cations and anions. In this parametrization of the Neue equation, the *S1* parameter reflects the difference between the logarithm of retention factors between the 0% organic modifier content and 100% organic modifier content.

Furthermore, a linear relationship between *pKa* values and organic modifier content was assumed:

(3)

where denotes dissociation constants of an analyte in given chromatographic conditions, denotes aqueous *pKa*, and denotes the slope due to changes in the organic modifier.

## Measurement-Error Model

The observed retention factors (*tRobs,z*) were modeled using the following model:

(4)

where *z* denotes the *zth* measurement and *student\_t* denotes the student’s t-distribution with the mean given by the predicted retention time *tR,z*, scale *σ* (analyte and column specific) and normality parameter . The retention time *tR,z* under an organic modifier gradient was calculated utilizing the well-known integral equation:

(5)

where *t0*denotes column hold-up (dead) time, *te* denotes extra column time, and denotes the instantaneous isocratic retention factor corresponding to the mobile phase composition at time *t* at the column inlet for a particular observation. The numerical solution of this integral equation was carried out using the method of steps with 4 and 10 steps for methanol and acetonitrile gradients using the method proposed by Nikitas et al.10

## Analyte-Level Model

Parameter were assumed to be different for each analyte to account for between analyte variability (BAV). The relationship was of the form:

(6)

where *Ri* denotes an individual value of a parameter *R*, is the mean value that correspond to an analyte with *logP=*2.2, is a regression coefficient, anddenotes standard deviation. For correlated parameter a multivariate normal distribution was used.The BAV was assigned for *logkw* and *S1m* (*S1* in MeOH) for neural forms of analytes (these parameters were correlated (ρ) and depended on *logP* (*β*)). Also the effects of ACN on *S1* (*dS1*), the effect of temperature on logkw (*dlogkT*), the effect of dissociation on *logkw*, *S1m* and *dS1* (*dlogkw*, *dS1m*, *ddS1)* separately for acids and bases), and the effect of column on these parameters (*clogkw*, *cS1m*, *cβ*, *cdS1*, *cdlogkT*, *cdlogkw*, *cdS1m*, *cddS1*) were used to characterized the impact of chromatographic conditions on model parameters. Also the *clogkw* for neutral forms of analytes was assumed to be corelated (cρ). The pH effects for cations and anions (*apH*) were assumed to be the same across analytes but different across the columns (*capH*). S2 parameter was assumed to be the same across analytes and columns but different in MeOH and ACN (*logS2m*, and *dlogS2* for ACN effect). Also the BAV was assigned for dissociation related parameters: *pKaw* (dependent on literature *pKa* values), the parameters for acids and bases in MeOH (, and the effect of ACN on (. The dissociation related parameters were assumed to be independent of the column. The standard deviation of BAV was denoted as ω for the parameters of the neutral form of analyte on XBridge Shield RP18 and cω for the difference between the other columns and XBridge Shield RP18, *ωT*and *cωT* for temperature effects, κ and cκ for the parameter related to dissociated forms and τ for the parameters related to dissociation constant.

## Priors

The Bayesian model requires specification of priors that allow to incorporate domain expertise into inferences. In this work priors were selected based on literature knowledge as discussed previously.4 The means for the priors describing column effects were assumed 0. The standard deviation were set to about half of the standard deviations/scales used for XBridge Shield RP18 parameters. It assumes a general similarity of the columns.

## Bayesian Inference

**Technical.** Multilevel modeling was performed in Stan/cmdstanr33 software linked with Rstudio34. For the inference we used 8 Markov chains with 500 iterations after 1000 warm up iterations. The *reduce\_sum* function was used to accelerate the calculations by parallelizing the execution of a single Stan chain across multiple cores. Convergence diagnostics were checked using Gelman-Rubin statistics and trace plots. No divergence was reported in the model. The R code, data and Stan code used to analyze the data are publicly available from GitHub (https://github.com/wiczling/columncomparison). The raw data are also available through a repository. The calculations were run on the Tryton computing cluster in Centre of Informatics Tricity Academic Supercomputer.

**Predictions Using a Limited Set of Experiments.** The model applicability was illustrated using selected 6 analytes with different acidic/basic properties: acridine (monoprotic acid), baclofen (zwitterion: acidic and basic group), hydrocortisone (neutral), pioglitazone (zwitterion: basic and acidic group), quinine (diprotic: 2 basic groups), and tolbutamide (monoprotic base). In this work we present the limited data predictions that correspond to a future predictions given all the experimental data collected for XBridge Shield RP18 column. For this problem all the population parameters were fixed to the final model estimates. The individual parameters for our six analytes were re-estimated using limited data. It allowed to assess the accuracy of prediction on other columns in a situation of having a very good understating of retention in one reference column (specifically XBridge Shield RP18 column).

**Uncertainty chromatogram**. The predictions were summarized as uncertainty chromatograms (posterior distribution of retention times expected for a given set of chromatographed analytes under given conditions).12 The uncertainty chromatogram visualizes the uncertainty for the locations of the maximum of each peak on a given chromatogram. Any area under the uncertainty chromatogram for a particular analyte can be probabilistically interpreted as a fraction of analytes (similar with respect to predictors and gathered data) that are expected to have a retention time within the range that the area was calculated.

**Decision making.** The Bayesian optimal decision was sought based on the maximum of expected utility calculated across a wide range of chromatographic conditions. The utility function (*U*) was specified to each possible chromatogram (posterior predicted retention time). It was defined based on the lowest retention time across analytes (*mintr*), the highest retention time across analytes (*maxtr*), and the difference in retention times between the critical pair of analytes (*res*). This utility was set to zero if at least one of the analytes has retention higher than 40 or less than 2 min, and if the *res* is less than 2. Otherwise, the utility was linearly related to maxtr.

(7)

The expected utility was plotted for a dense grid of chromatographic conditions (the expected utility map). Based on that graph one can identify regions of chromatographic conditions leading to the desired/optimal separation given access to various data.

# RESULTS AND DISCUSSION

In this work, we applied a mechanistic model to describe the retention data of small molecules obtained for a wide range of chromatographic conditions and for five chromatographic columns. The model was built based on known fundamentals of gradient chromatography and prior knowledge available in the literature. The methodology allowed us to characterize the stationary phases properties using the commonly used parameters in chromatography that allow for an easy interpretation and application of this parameters in decision making. The model parameters are summarized in Table S1 and Figure S2. The key parameters are also provided in Figures 1-3. These parameters provide a concise summary of influence of various factors affecting retention of analytes on the tested columns. Basically, each parameter isolates and quantitates an effect of a chromatographic condition/analyte properties (pH, organic modifier type and content, temperature, column type, log P) on the retention of neutral, acidic and basic form of an analyte. As an example, the typical logkw of a neutral form of an analyte (a measure of hydrophobicity) is 3.6 for XBridge Shield RP18, and is 0.42, 0.17, 0.10, and 0.17 higher for XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8 columns. The typical slope in MeOH is 4.96 and the difference between XBridge Shield RP18 and the other columns is 0.59, -0.12, 0.36, 0.48. The BAV for logkw and S1 (ωlogkw and ωS1) for XBridge Shield RP18 is about 0.9 with a high correlation (ρ=0.87). The BAV for column effects on logkw (cωlogkw) are small (0.10-0.13). They are also correlated with correlations (cρ) ranging from 0.55-0.92 depending on the compared columns. The BAV for column effects on S1 in MeOH (cωS1m) are also small (0.02-0.15). Since the between analyte variabilities are on a log (with base 10) scale, all standard deviations has to be interpreted accordingly for the retention factor. For example the between analyte variability of 0.1 for clogkw implies that the predicted effects of the column are on the order of ± 0.1, which corresponds to a multiplicative factors from exp(-2.3\*0.1)=0.79 to exp(2.3\*0.1)=1.26.

The log P effect for logkw and S1m have slope of 0.83 and 0.48 for XBridge Shield RP 18. The effects of columns on this parameter are small. The largest difference is between XBridge C8 and Xterra MS C18 (about 0.18) for cS1.

The S1 is higher in ACN than in MeOH by 0.61 for XBridge Shield RP18 column, and this difference is higher by 0.15 0.81, 0.51, 0.04 for XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8. The BAV for dS1 (ωdS1) is about (0.55) and for column effect range from (0.14-0.47).

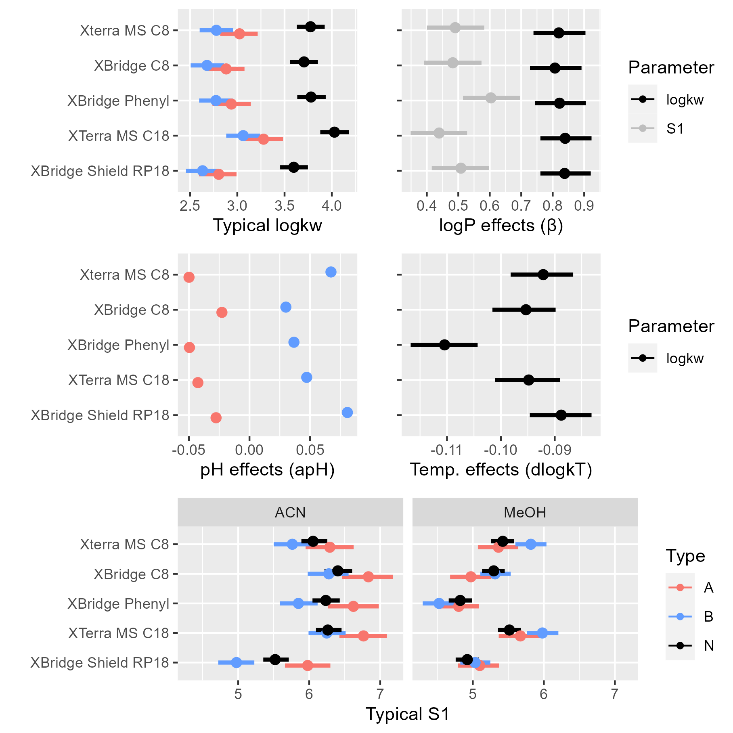


Figure 1. Summary of marginal posterior distributions of the key population-level parameters characterizing retention of analytes in the tested columns.

The effect of temperatures are small and consisted with prior knowledge. The column effects on temperature seems to the larges for XBridge Phenyl column.

The logkw for acids and bases is lower by -0.79 and -0.97 on XBridge Shield RP18. The column effects for that parameter are very small. The pH effects on logkw for acids and bases (apH) are small and negative for acids and positive for bases. This effects reflects the changes in stationary phase properties due to pH of the mobile phases.

…

The model predictions are well calibrated with the data, as shown in Figure S3. The individual and population predictions versus observed retention times are relatively symmetrically distributed around the line of identity indicating model applicability for predictions. The individual and population predictions for several analytes are shown in Figure S4 and S5. The individual prediction are very precise and close to observed data. The population predictions are also well calibrated but are less precise. The limited data predictions are shown in Figure S6. By comparing them to the population predictions one is able to assess the added predicted value of XBridge Shield RP18 data. Uncertainty chromatograms for one chromatographic conditions are shown in Figures S7.

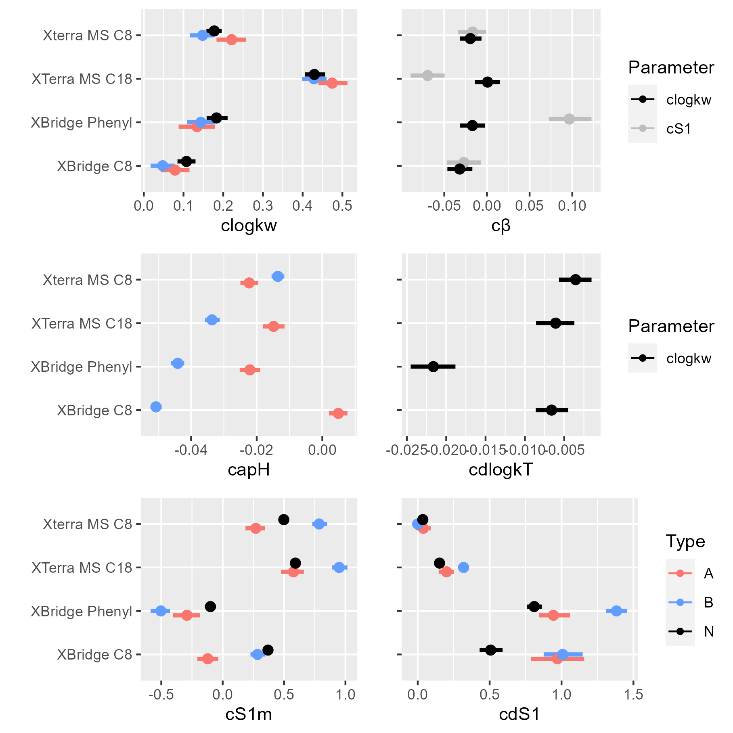


Figure 2. Summary of marginal posterior distributions of the key population-level parameters characterizing the difference in retention of analytes between the indicated column and XBridge Shield RP18 column (column effects).

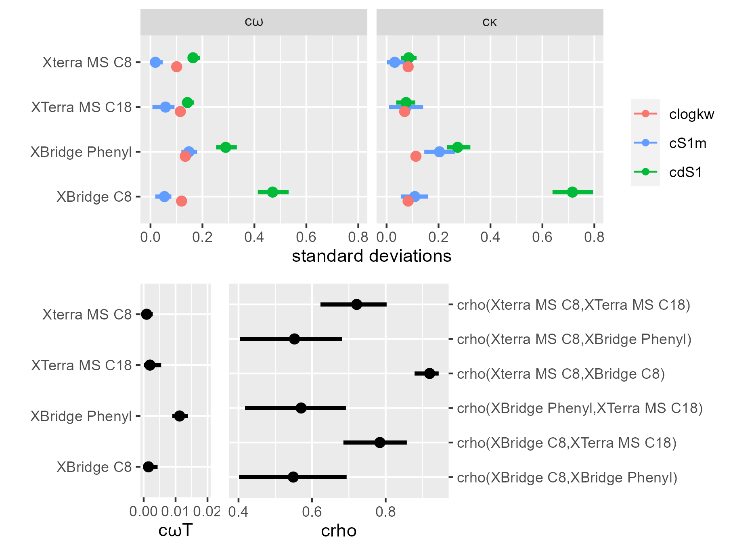


Figure 3. Summary of marginal posterior distributions of the BAV parameters characterizing the difference in retention of analytes between the indicated and reference XBridge Shield RP18 column.

All model parameters jointly affect analyte retention. To better illustrate the joined effect of parameters, we simulated the retention factors for the typical acidic, basic and neutral analyte with log P of 0, 3 and 6. The results are present in Figure 4. We also provide various isocratic prediction for 6 selected analytes to illustrate the impact of parameters on isocratic retention (Figure S8 and S9).

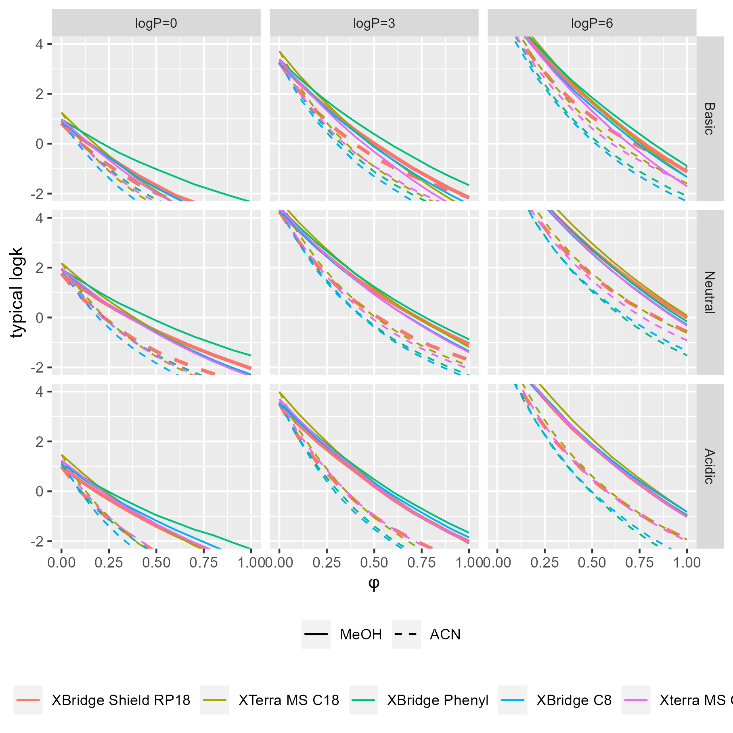


Figure 4. Graphical display of the typical retention profiles (log k vs φ) for basic, neutral and acidic form of an analyte with log P of 0, 3, and 6 at 25oC. The uncertainty is not present to improve readability.

The population predictions are very uncertain and of limited usefulness. Basically, they are driven by unexplained BAV *ω*, which is large. However, by adding experimental information one can decrease this uncertainty. For example, including all the experimental data provides a very accurate predictions (individual predictions). For the limited data predictions the uncertainty is almost entirely reduced for XBridge Shield RP18 column (as there is a lot of data collect using that particular column). For other columns there is still some proportion of uncertainty left, here mostly driven by c*ω*. Since cωis small (about 0.1 for neutral forms of analyte)one can expect a fairly precise predictions, i.e. one is able to predict isocratic retention factor with an uncertainty of about 20-25%. It might be of practical usefulness for problems involving few analytes.

The model can be used in decision making to help identify chromatographic conditions leading to the desired separation. In this work it was illustrated by specifying a very simply utility function favoring shorter runs within a separation window between 2 and 40 min and ensuring at least 2 min difference in retention of critical pairs of analytes. The utility maps are presented on Figures S10 and S11 and the uncertainty chromatograms for the highest expected utility are shown on Figures 5 and 6 for individual and limited data predictions. Based on individual predictions one can identify chromatographic conditions leading to reasonable chromatograms for all of the columns. The highest utility can be expected for Xterra MS C8. Although the differences are small. There are also several “windows” of opportunity presented on the

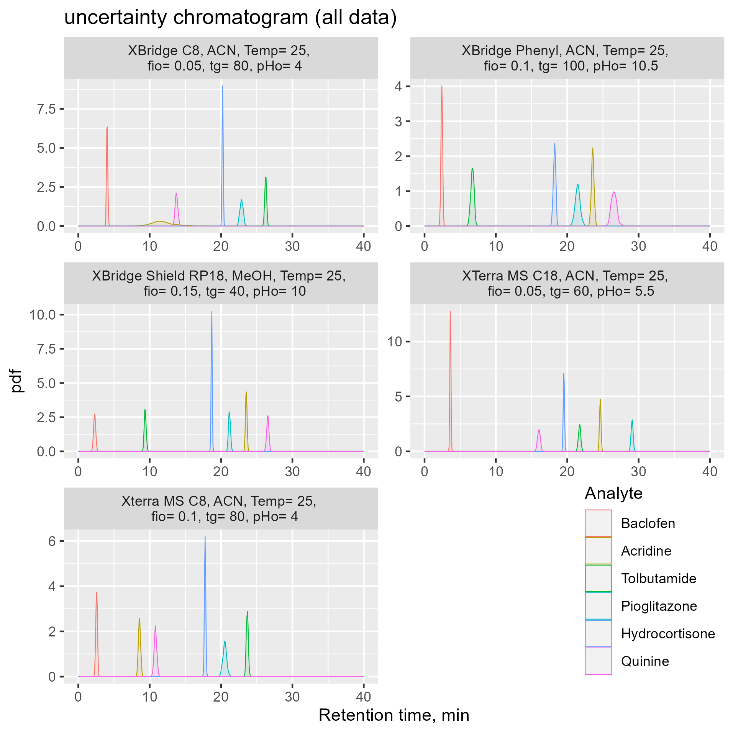


Figure 5. The optimal Bayesian-based chromatographic conditions found based on individual data predictions.

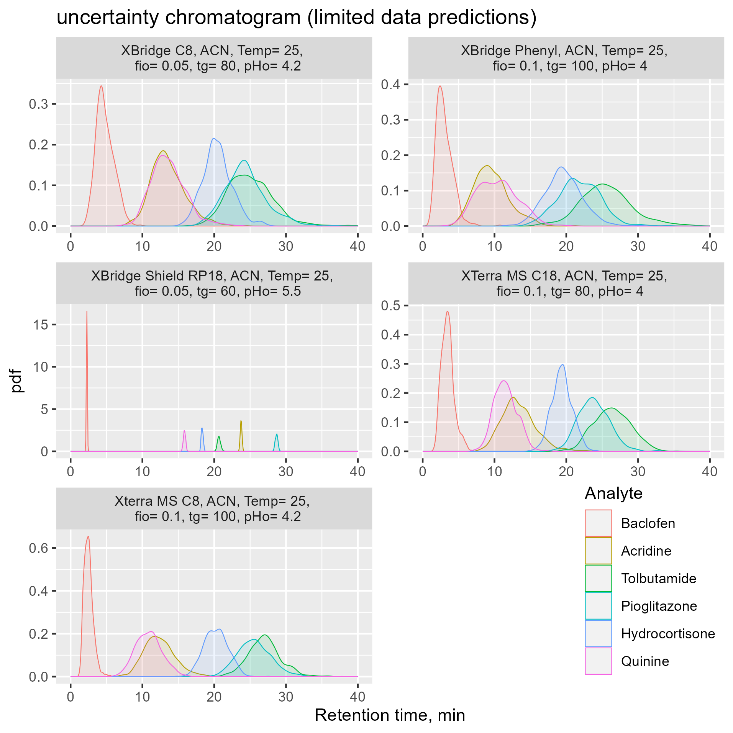


Figure 6. The optimal Bayesian-based chromatographic conditions found based on limited data predictions.

The proposed model is complex but there still are some uncounted complexities that can be added (e.g. temperature effects for the dissociated forms, BAV for S2 and apH, etc.). Specifically, the S2 parameter was assumed similar across columns. This is a strong assumptions, but justified by the fact that any small change of S2 can be equally well accounted for by changes in S1. We also encountered several technical difficulties that required several simplifications to run the MCMC without problems, e.g. excluding analytes with more than 2 dissociations steps.

There are possible several modification of the proposed model. For example one could include column as a random effects and add column-level predictors. It would allow to build a model that generalized to other columns in a similar way the proposed model generalized to other analytes. This idea is worth pursuing, however requires data collected for a wider range of columns.

One possible application of the model is in finding the best experimental design ensuring the highest informativeness of a set of preliminary (scouting) experiments. Since our model provides a prior information for subsequent analysis this prior information can be utilized for that purpose to decrease the required number of experiments.

The results lead to the conclusion that it is rather hopeless to predict analyte retention precisely without a set of preliminary chromatographic measurements. Even having access to the extensive data collected for one column, the retention time can be predicted with ~25% precisions for the most optimistic scenario of a neutral compound. It might by sufficient for some problems but is not sufficient in general.

# CONCLUSIONS

This work demonstrates the application of a Bayesian multilevel model to compare various stationary phases using large datasets collected for a wide range of chromatographic conditions. This analysis characterizes the chromatographic retention of neutral, acidic, and basic analytes. It also provides a way to characterize the influence of pH, temperature, organic modifier type and content on analyte retention across a range of columns.

# ASSOCIATED CONTENT

## Supporting Information

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

# 1. Model description; 2. Table S1. Summary of the MCMC simulations of the marginal posterior distributions of population-level model parameters; 3. Figure S1. Raw data for 6 selected analytes; 4. Figure S2. Summary of the MCMC simulations of the marginal posterior distributions of population-level model parameters; 5. Figure S3. Goodness of fit plots; 6. Figure S4. Individual gradient predictions; 7. Figure S5. Population gradient predictions; 8. Figure S6. Limited data gradient predictions; 9 Figure S7. Uncertainty chromatograms. 10. Figure S8. Individual isocratic predictions; 11. Figure S9. Population isocratic predictions; 12. Figure S10. Utility maps based on individual predictions; 13. Figure S11. Utility maps based on limited data predictions.

# AUTHOR INFORMATION

## Corresponding Author

\* wiczling@gumed.edu.pl

## Author Contributions

ŁK, JJ, WSL collected the experimental data; ŁK and AK prepared the data for analysis; AK and PW analyzed the data; PW, AK wrote the paper with input from all authors; PW conceived of the presented idea, designed the study and supervised the project; and MM helped supervise the project.

Notes  
The authors declare no competing financial interest.

# FUNDING

This project was supported by the National Science Centre, Poland (grant 2015/18/E/ST4/00449). AK was also supported by the project POWR.03.02.00-00-I035/16-00 cofinanced by the European Union through the European Social Fund under the Operational Programme Knowledge Education Development 2014–2020. Calculations were carried out at the Academic Computer Centre in Gdansk

# REFERENCES

(1) Žuvela, P.; Skoczylas, M.; Jay Liu, J.; Ba̧czek, T.; Kaliszan, R.; Wong, M. W.; Buszewski, B. Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chem. Rev.* **2019**, *119* (6), 3674–3729. https://doi.org/10.1021/acs.chemrev.8b00246.

(2) Gritti, F.; Guiochon, G. Adsorption Mechanism in RPLC. Effect of the Nature of the Organic Modifier. *Anal Chem* **2005**, *77* (13), 4257–4272. https://doi.org/10.1021/ac0580058.

(3) 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.

(4) 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.

(5) Kamedulska, A.; Kubik, Ł.; Wiczling, P. Statistical Analysis of Isocratic Chromatographic Data Using Bayesian Modeling. *Anal Bioanal Chem* **2022**, *414* (11), 3471–3481. https://doi.org/10.1007/s00216-022-03968-x.

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

(7) Wiczling, P. Analyzing Chromatographic Data Using Multilevel Modeling. *ANALYTICAL AND BIOANALYTICAL CHEMISTRY*, 2018, *410*, 3905–3915. https://doi.org/10.1007/s00216-018-1061-3.

(8) Haider, N. Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: An Open-Source Approach. *Molecules* **2010**, *15* (8), 5079–5092. https://doi.org/10.3390/molecules15085079.

(9) ACD/Labs. *Release 12.0*; Advanced Chemistry Development Inc.: Toronto, ON, Canada, www.acdlabs.com, 2022., 2011.

(10) Nikitas, P.; Pappa-Louisi, A. New Equations Describing the Combined Effect of PH and Organic Modifier Concentration on the Retention in Reversed-Phase Liquid Chromatography. *J Chromatogr A* **2002**, *971* (1–2), 47–60. https://doi.org/10.1016/s0021-9673(02)00965-2.

(11) 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.

(12) Jano, I.; Hardcastle, J. E.; Zhao, K.; Vermillion-Salsbury, R. General Equation for Calculating the Dissociation Constants of Polyprotic Acids and Bases from Measured Retention Factors in High-Performance Liquid Chromatography. *J Chromatogr A* **1997**, *762* (1–2), 63–72. https://doi.org/10.1016/s0021-9673(96)00739-x.

TABLE OF CONTENTS (TOC)

