Comparison of chromatographic stationary phases using Bayesian-based multilevel modeling

Paweł Wiczling\*, Agnieszka Kamedulska

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 an initial 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 chromatographic parameters of neutral, acidic, and basic analytes. The analysis 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 proposed method seems to be an interesting alternative to existing approaches used to compare chromatographic stationary phases.

The selection of stationary phases is an important aspect of analytical work. Several methods and approaches has been proposed over the years to classify, select, and characterize chromatographic stationary phases. They has been recently reviewed in literature.1,2 The chromatographic stationary phases are usually characterized using probe analytes and a fairly limited set of chromatographic conditions, aiming to estimate the most important chromatographic characteristics, e.g. column efficiency, hydrophobicity, silanol activity, ion-exchange capacity, steric selectivity and the amount of metal impurities. Usually results are presented without building a formal statistical model that limits the practical usefulness of these methods in extrapolating the obtained results to other analytes and other chromatographic conditions. Also a very detailed physical (and thermodynamic) description of the chromatographic systems is possible, nevertheless it requires an extensive sets of carefully controlled experiments and as such are limited to few analytes.3

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 based on various number of preliminary experiments (e.g. to predict retention time for a set of analytes given several measurements collected using a different stationary phase). The accurate predictions of the retention time across a range of columns in liquid chromatography is required or rapid column screening, computer-assisted method development and method transfer.4

In this work we applied the previously developed Bayesian multilevel framework5–8 to characterize chromatographic gradient retention time datasets collected using a multicomponent mixtures of analytes, five stationary phases (XBridge Shield RP18, XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8), and a wide range of chromatographic conditions (pH, organic modifier, temperature, gradient program). The general idea is to use a mechanistic model of chromatographic retention that jointly characterize the retention of available data, allow to estimate the column effects on chromatographic parameters for acids, bases and neutral analytes, quantify various sources of variation, and characterizes uncertainty. The usefulness of the model and results in decision making will be illustrated by searching chromatographic conditions leading to the desired separation given access to different types of preliminary data, e.g. predicting retention on the tested columns given XBridge Shield RP18 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).9 The lipophilicity (*log P*), dissociation constant (*pKalit*) were calculated using the ACD/Labs program10 based on the structures of analytes generated from SMILES strings.

More details about data extraction can be found in our previous work.5 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.11,12 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* forms13:

(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 with the base of 10 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.11

## 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 *logkwN* and *S1mN* (*S1* in MeOH) for neural (N) form of analytes (these parameters were correlated (ρ) and depended on *logP* (*β*)). Also the effects of ACN on *S1* (*dS1N*), the effect of temperature on logkw (*dlogkT*), the effect of dissociation on *logkw*, *S1m* and *dS1* (*dlogkw*, *dS1m*, *ddS1)* separately for acids (A) and bases (B)), 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 correlated (cρ). The pH effects for cations and anions (*apHA* and *apHB*) were assumed to be the same across analytes but different across the columns (*capHA* and *cpHB*). 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 *pKa* and 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.5 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/cmdstanr14 software linked with Rstudio15. 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 repository (https://github.com/wiczling/columncomparison). The raw data are also available through a osf.io repository16–20. 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 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.

**Decision making.** The Bayesian optimal decision was sought based on the maximum of expected utility calculated across a wide range of chromatographic conditions. Simulations were performed for *fio* ranging from 0.05 to 0.15 by 0.02, *pHo* ranging from 2.5 to 10.5 by 0.2, *tg* ranging from 20 to 200 by 20, using MeOH and ACN as an organic modifiers and for five tested columns at 25oC. The utility function (*U*) was specified to each possible chromatogram (posterior predicted retention time). It was defined based on the lowest retention time among analytes (*mintr*), the highest retention time among analytes (*maxtr*), and the difference in retention times between the critical pair of analytes (*res*). This utility was assumed 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* favoring shorter runs.

U (7)

The expected utility was plotted for a dense grid of chromatographic conditions yielding the expected utility map. The graph was used to identify chromatographic conditions leading to the desired/optimal separation.

# 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 chromatographic parameters (specifically *logkw*, *S1* and *S2* of the Neue model). It also allows for an easy interpretation of the estimated effects and application of results in decision making. The model parameters are summarized in Table S1 and Figure S2. The key parameters are also provided in Figures 1-3.

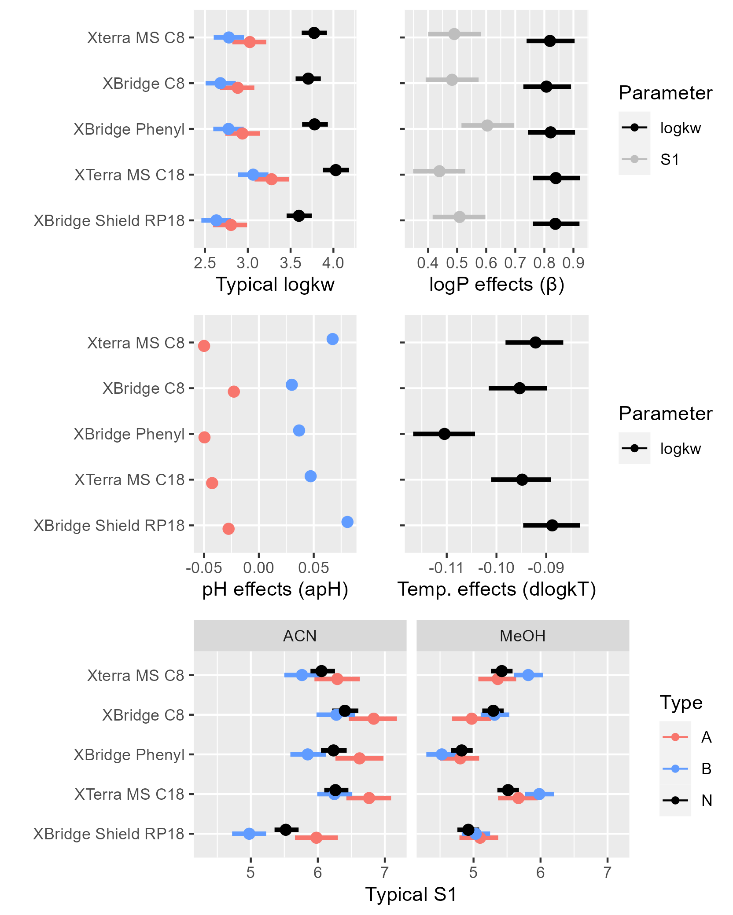


Figure 1. Summary of marginal posterior distributions of the key population-level parameters characterizing retention of analytes for all the tested columns.

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. Here to simplify the text we will present the mean values of posterior distribution. The whole posterior distribution is presented graphically in the attached figures and tables. As an example, the typical *logkw* of a neutral form of an analyte (a measure of hydrophobicity) is 3.6 for XBridge Shield RP18 at 25oC, 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. It corresponds to a typical 2.64 1.48, 1.26 1.48-fold larger retention factors for a neutral form of an analyte in neat water eluents in comparison to XBridge Shield RP18. 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 of 0.87 implying large mutual information of the analyte specific values of these two parameters. 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 logarithmic scale with base 10, all standard deviations has to be interpreted accordingly to understand the variability of retention factor. For example the between analyte variability of about 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 for retention factor ranging from 10-0.1=0.79 to 100.1=1.26.

The log P effect for *logkw* and *S1m* has a slope of 0.83 and 0.48 for XBridge Shield RP 18. The column effects on this parameter are small. The largest effect of 0.1 was observed for XBridge Phenyl *cS1m* parameter.

The *S1* is higher in ACN than in MeOH by 0.61 for XBridge Shield RP18 column. In addition it is higher by 0.15, 0.81, 0.51, 0.04 for the toher columns XTerra MS C18, XBridge Phenyl, XBridge C8 and Xterra MS C8. The BAV for dS1 (ωdS1) is about (0.55) and for column effects it ranges from (0.14-0.47).The S2 parameter of the Neue equation is 0.49 in MeOH and 1.3 in ACN. The temperature effect for XBridge Shield RP18 is small and consistent with prior knowledge (-0.09 per increase in 10oC) with BAV of 0.03. The column effects on temperature and corresponding BAV are small. The largest effects of -0.02 was observed for XBridge Phenyl.

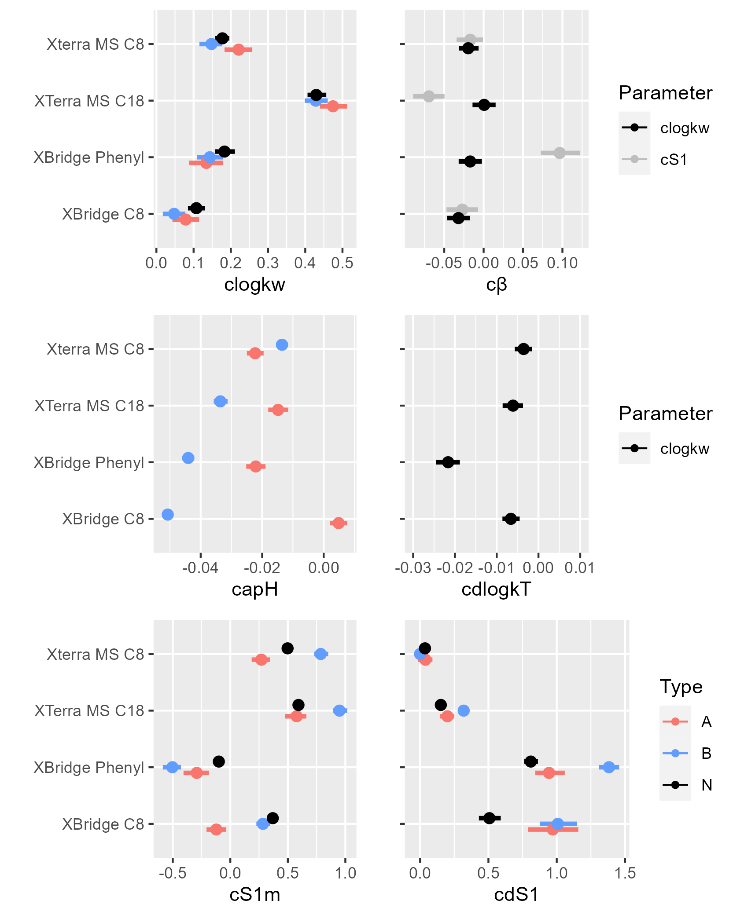


Figure 2. Summary of marginal posterior distributions of the key population-level parameters characterizing the difference in retention of analytes between the indicated column and XBridge Shield RP18 column (column effects).

The logkw for acids and bases is lower by -0.79 and -0.97 on XBridge Shield RP18 with BAV of 0.59. The column effects for that parameter are very small (from -0.04 to 0.04) with BAV ranging from 0.07-0.11. The pH effects on logkw for acids and bases (apH) are small and negative for acids (-0.03 per unit pH) and positive for bases (0.08 per unit pH) for XBridge Shield RP18. The column effects for *apH* vary across columns and range for acids from -0.02 to 0 and for bases from 0.05 to -0.01.

The S1m for acids and bases is higher by 0.17 and 0.12 on XBridge Shield RP18 with BAV of 0.69. The column effects for that parameter very from -0.49 to 0.36 depending on column an analyte form with BAV ranging from 0.04-0.20. The slope for acids and bases is different in ACN (relative to S1m) by 0.28 for acids and -0.67 for bases on XBridge Shield RP18 with BAV of 0.55. The column effects for that parameter very from -0.04 to 0.57 depending on column an analyte form with BAV ranging from 0.07-0.72 (the highest value was observed for XBridge C8).

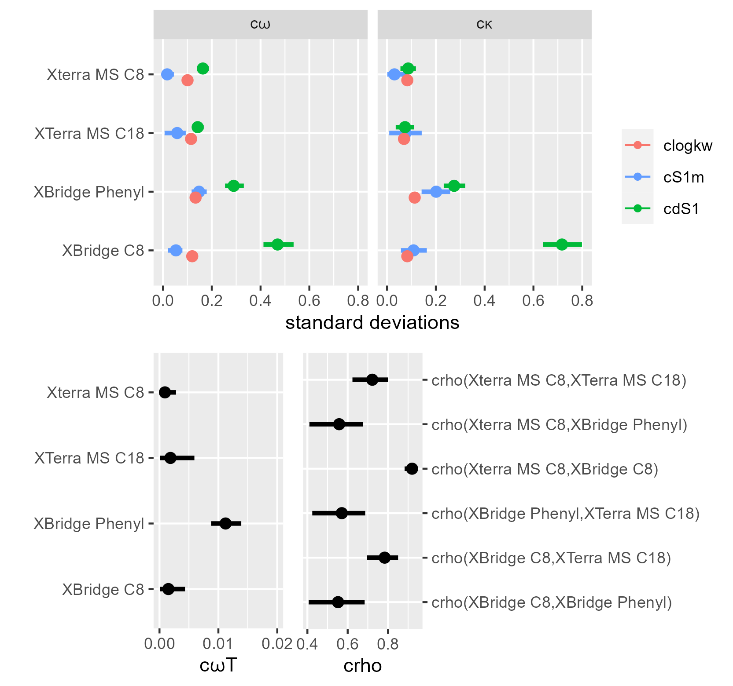


Figure 3. Summary of marginal posterior distributions of the BAV parameters characterizing the difference in retention of analytes between the indicated and reference XBridge Shield RP18 column.

The model predictions are well calibrated with the data for all the columns 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 also 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 as expected 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 population, individual and limited data predictions under the same chromatographic conditions are shown in Figure S7. They illustrate retention uncertainties given access to different data. Clearly the added predictive value of XBridge Shield RP18 data is large. As an example the standard deviation of retention time predictions is reduced from about 10 min for population predictions to about 2 min for limited data predictions and further to about 0.1 for individual predictions. It confirms that the population predictions are very uncertain and of limited practical usefulness. The reason is that 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 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 collected using that particular column). For all the other columns there is still some proportion of uncertainty left. In this case, it mostly depends on 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 simple problems involving few analytes.

All model parameters jointly affect analyte retention. To better illustrate this 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. Based on that graph one is able to directly compare column characteristics across a wide range of chromatographic conditions. We also provide various isocratic prediction for 6 selected analytes to illustrate the impact of parameters on isocratic retention (Figure S8 and S9).

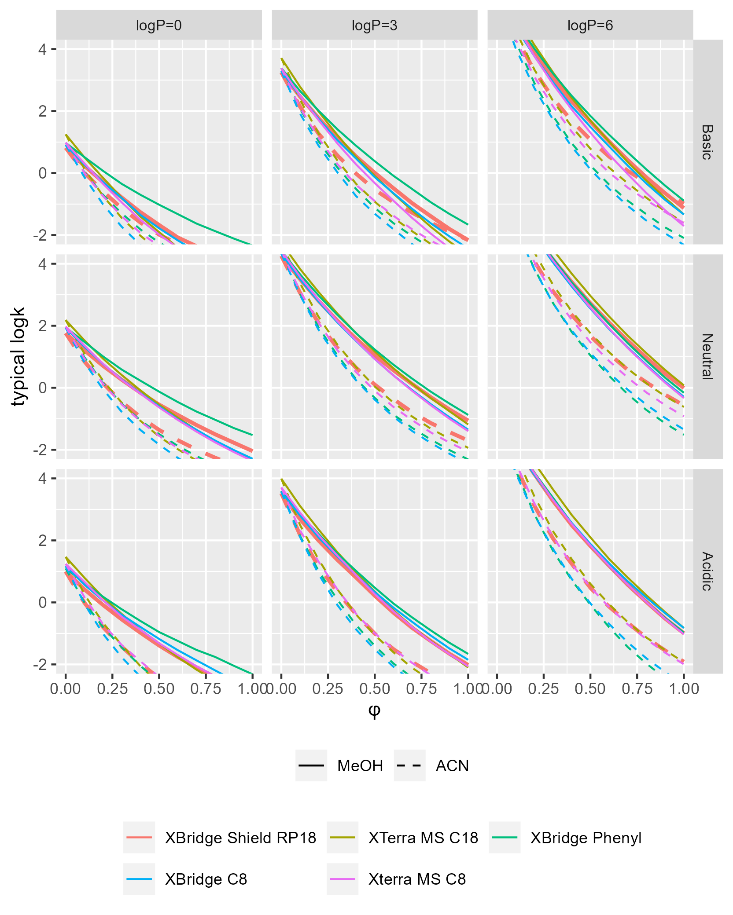


Figure 4. Graphical display of the typical retention profiles (log k vs φ) for a typical basic, neutral and acidic form of an analyte with log P of 0, 3, and 6 at 25oC. The uncertainty is not provided to improve readability.

The model can be used in decision making to help identify chromatographic conditions leading to the desired separation. In this work it was illustrated using statistical decision theory. For that purpose we proposed a very simply utility function that favors 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 large expected utilities for all of the analyzed columns. The highest expected utility among the simulated chromatographic conditions was identified for Xterra Phenyl. Nevertheless, the differences in expected utility are small suggesting that similar separations (with respect to the proposed utility) can be identified also for other columns. Based on utility maps one can also identify the “windows of opportunity” understood as a regions of chromatographic conditions leading to similar separations.

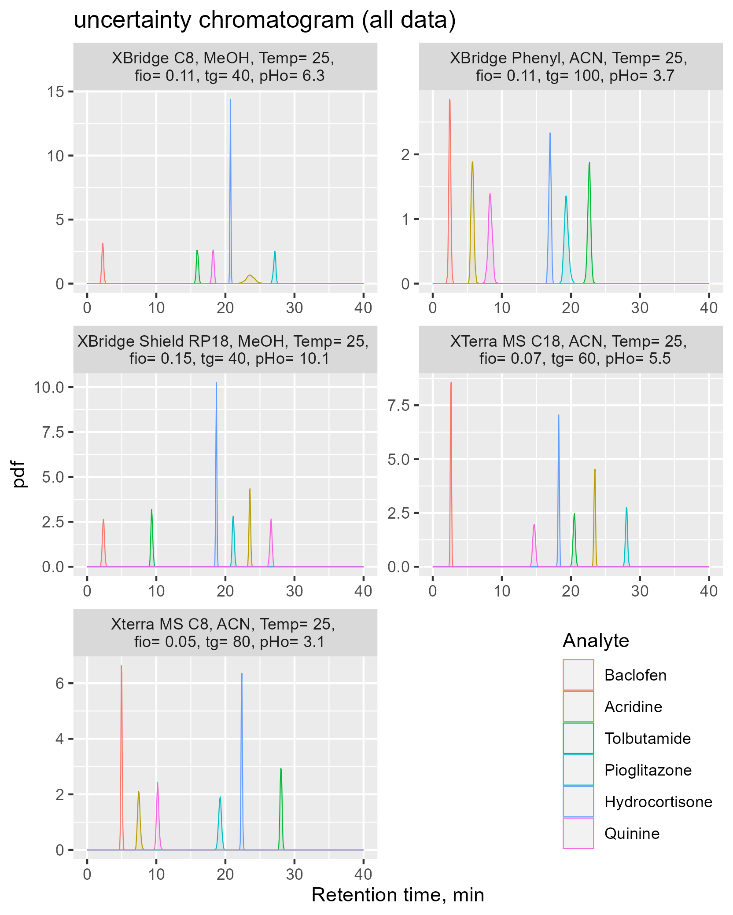


Figure 5. The optimal Bayesian-based chromatographic conditions found based on individual data predictions.

It is also possible to identify optimal conditions based on the limited data (e.g. conditional on XBridge Shield RP18 data). This is a situation of predicting retention on a different column having a fairly good understanding of retention of analytes for one reference column. In our scenario, the uncertainties of the resulting chromatograms are fairly small (standard deviation is in a range of few minutes for gradient conditions), however are not sufficient to identify the best conditions with confidence. Clearly some additional experiments are required to improve precision. There are several options of how to proceed. One can try to find the best experimental design ensuring the highest informativeness of a set of additional experiments. Since the proposed model provides a priori information for subsequent analysis this prior information can be utilized to decrease the number of required experiments. One can also perform experiments sequentially and refine predictions after every experiment.21 The clear benefit of having the model that quantifies uncertainty is that it allows to identify chromatographic conditions for which the desired separation is plausible.

The proposed model is complex, but there are still several improvements that can be considered (e.g. temperature effects for the effects of dissociation and *pKa*, BAV for S2 and apH, etc.). Specifically, the S2 parameters 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 during the model building process that required simplifications to run the MCMC without any convergence issues, e.g. excluding analytes with more than 2 dissociations steps and analytes which were incorrectly identified (there was a mismatch between observed pH-profile and ACD dissociation pattern). In the present model the column effect were incorporated as fixed effects. It is also possible to include them 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. In this work we decided to use log P as the only predictor relating molecular structure to chromatographic parameters. In principle various functional groups can be added that would allow to assess the impact of molecular structure on retention in a more nuanced fashion7. It would lead to a more detailed characterization of column properties. This again required a dataset with larger number of structurally diverse analytes.

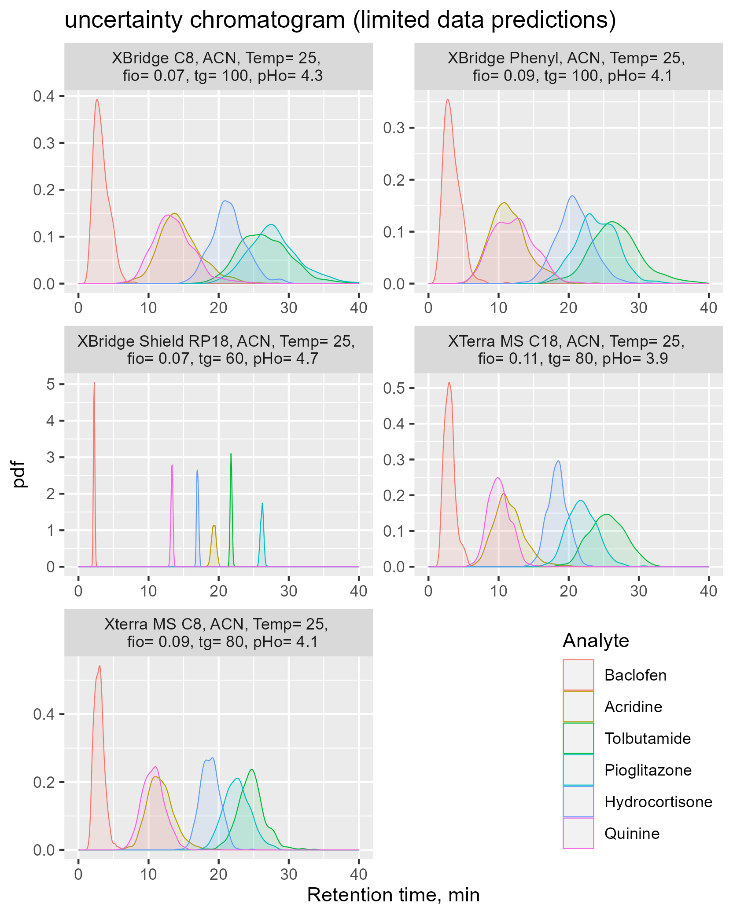


Figure 6. The optimal Bayesian-based chromatographic conditions found based on limited data predictions.

Our results confirm that it is rather difficult to predict analyte retention precisely without a set of preliminary chromatographic measurements. Even, having access to the extensive data collected for one column, the isocratic retention factor can be predicted with ~25% precisions for the most optimistic scenario of a neutral compound. It might by sufficient for some simple problems encountered in practice, but is not sufficient for more complex problems involving several analytes of different acidic/basic characteristics. We show that using a fairly difficult problem of finding the optimal separation for 6 diverse analytes. It is also rather difficult to provide a very simple and general rule of improving selectivity based on the model. The preferred way is to use the model-based simulations and user-specific decision functions to find the conditions leading to the desired separation.

<the currently used methods of column classification are usually based on a similarity measure (usually without an any notion of uncertainty). In our opinion this can be improved by reporting similarity measures as a function of a chromatographic conditions that are also specific for a given problem at hand. Basically it is rather hard to find a single number to characterize various interactions occurring in the chromatographic system. We think it is more valuable to provide a model quantitating the variability and uncertainty of relevant chromatographic parameters, that is sufficiently complex to generalize to other chromatographic conditions, analytes and chromatographic columns. Such a model might be built once for a particular stationary phase and used by others to solve their specific problem.>

<We want to emphasize here that because population predictions are very uncertain also the column ranking without any experimental data cab be expected to be very uncertain. One needs to perform few measurements using at least one column to have practically useful results. >

<fixed effects are adjustable. Based on them one can change chromatographic conditions to adjust for the typical differences. The unexplained BAV controls how likely it is for a set of analytes to have a similar retention>

# CONCLUSIONS

This work demonstrates the application of a Bayesian multilevel model to compare five stationary phases using large datasets of retention times collected for a wide range of gradient chromatographic conditions. This analysis characterizes the chromatographic retention of neutral, acidic, and basic analytes and assesses the effects of pH, temperature, log P, *pKa* organic modifier type and content on analyte retention across a range of columns. The model along with the estimated parameters can be used to solve various decision problems. The proposed method provides a tool allowing for a very comprehensive assessment of stationary phase properties that seems to be an interesting alternative to existing methods.

# 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

PW and AK analyzed the data; PW wrote the paper; PW conceived of the presented idea, designed the study and supervised 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

# ACKNOWLEDGEMENT

I wish to acknowledge the help provided by Łukasz Kubik, Julia Jacyna, Wiktoria Struck-Lewicka, Michał J. Markuszewski in collecting the experimental data.

# 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) Visky, D. 4 Column Characterization and Selection. In *HPLC Method Development for Pharmaceuticals*; Ahuja, S., Rasmussen, H., Eds.; Separation Science and Technology; Academic Press, 2007; Vol. 8, pp 85–109. https://doi.org/10.1016/S0149-6395(07)80010-9.

(3) 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.

(4) 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.

(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) 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.

(7) 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.

(8) 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.

(9) 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.

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

(11) 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.

(12) 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.

(13) 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.

(14) *Stan Development Team. YEAR. Stan Modeling Language Users Guide and Reference Manual, VERSION. https://mc-stan.org*. stan-dev.github.io. //mc-stan.org/users/citations/ (accessed 2022-01-18).

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

(16) Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M. J.; Wiczling, P. LC-TOF-MS Data Collected for 300 Small Molecules. XBridge Shield RP18 Column. *osf.io/* **2022**, *1* (1), 1. https://doi.org/10.17605/OSF.IO/ZQTJ7.

(17) Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M. J.; Wiczling, P. LC-TOF-MS Data Collected for 300 Small Molecules. XTerra MS C18 Column. *osf.io/* **2022**, *1* (1), 1. https://doi.org/10.17605/OSF.IO/QBV7J.

(18) Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M. J.; Wiczling, P. LC-TOF-MS Data Collected for 300 Small Molecules. XTerra-C8 Column. *osf.io/* **2022**, *1* (1), 1. https://doi.org/10.17605/OSF.IO/2MCNW.

(19) Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M. J.; Wiczling, P. LC-TOF-MS Data Collected for 300 Small Molecules. XBridge Phenyl Column. *osf.io/* **2022**, *1* (1), 1. https://doi.org/10.17605/OSF.IO/EVUJ9.

(20) Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M. J.; Wiczling, P. LC-TOF-MS Data Collected for 300 Small Molecules. XBridge-C8 Column. *osf.io/* **2022**, *1* (1), 1. https://doi.org/10.17605/OSF.IO/Y6S8P.

(21) 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.

TABLE OF CONTENTS (TOC)

