analytical chemistry

pubs.acs.org/ac Article

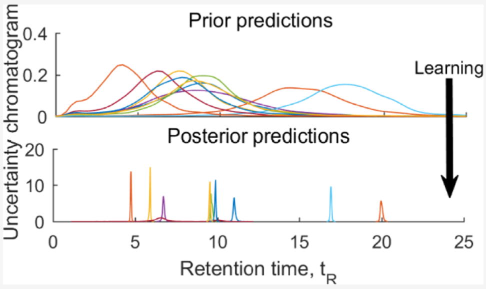
Application of Bayesian Multilevel Modeling in the Quantitative Structure–Retention Relationship Studies of Heterogeneous Compounds

Paweł Wiczling,* Agnieszka Kamedulska, and Łukasz Kubik

Cite This: https://doi.org/10.1021/acs.analchem.0c05227 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Quantitative structure–retention relationships (QSRRs) are used in the field of chromatography to model the relationship between an analyte structure and chromatographic retention. Such models are typically difficult to build and validate for heterogeneous compounds because of their many descriptors and relatively limited analyte-specific data. In this study, a Bayesian multilevel model is proposed to characterize the isocratic retention time data collected for 1026 heterogeneous analytes. The QSRR considers the effects of the molecular mass and 100 functional groups (substituents) on analyte-specific chromatographic parameters of the Neue model (i.e., the retention factor in water, the retention factor in acetonitrile, and the curvature coefficient). A Bayesian multilevel regression model was used to smooth noisy parameter estimates with too few data and to consider the uncertainties in the model parameters. We discuss the benefits of the Bayesian multilevel model (i) to understand chromatographic data, (ii) to quantify the effect of functional groups on chromatographic retention, and (iii) to predict analyte retention based on various types of preliminary data. The uncertainty of isocratic and gradient predictions was visualized using uncertainty chromatograms and discussed in terms of usefulness in decision making. We think that this method will provide the most benefit in providing a unified scheme for analyzing large chromatographic databases and assessing the impact of functional groups and other descriptors on analyte retention.



Quantitative structure–retention relationships (QSRRs) are used in the field of chromatography to model the relationship between an analyte structure reflected by various descriptors and chromatographic retention.[1] One of the purposes of building a QSRR is to predict analyte retention based on the chemical structure. It is occasionally called a "Holy Grail" problem in chromatography.[2] However, the accuracy of such predictions is often poor for analytes with known structures, as explained by Snyder et al.[2] "In general, it has not been proven possible to predict chromatographic retention in high-performance liquid chromatography (HPLC) with an accuracy that is anywhere near sufficient to support method development". Kaliszan[1] presents a contrary view: "because of QSRR, an optimization of chromatographic conditions can rationally be guided to provide a good separation of a given structurally defined analyte". At this point, it is critical to highlight that chromatographic models with a QSRR are built to predict analyte retention with access to a limited number of preliminary data, usually without any experimental data, and require decisions to be made under uncertainty; thus, these uncertainties should be quantitated and well calibrated. To make decisions with limited data, the analyst must consider a plausible range of chromatograms expected, given the available knowledge (i.e., analyte structure and/or experimental data) to make rational decisions. Bayesian methods are particularly well suited to show all necessary input to make decisions under uncertainty.[3−6]

There are many methods and approaches used in the field of chromatography and chemometrics to obtain QSRR models. These methods are reviewed in detail elsewhere.[1,7−9] In brief, QSRR models are typically built by considering fairly small data sets (i.e., considering congeneric compounds) and considering relationships between a single chromatographic parameter (i.e., retention in neat water as an eluent, log k_w) and a large set of structural descriptors. The applicability domain of such equations is often limited, if even specified. Frequently, multiple linear regression is used to estimate QSRR regression parameters. However, model building when there are many descriptors quantifying chemical structure and a relatively small number of analytes is particularly difficult and requires regularization to obtain useful results. Some challenges can be resolved using various chemometrics/machine learning techniques, including rank support vector machines, support

Received: December 14, 2020
Accepted: March 23, 2021

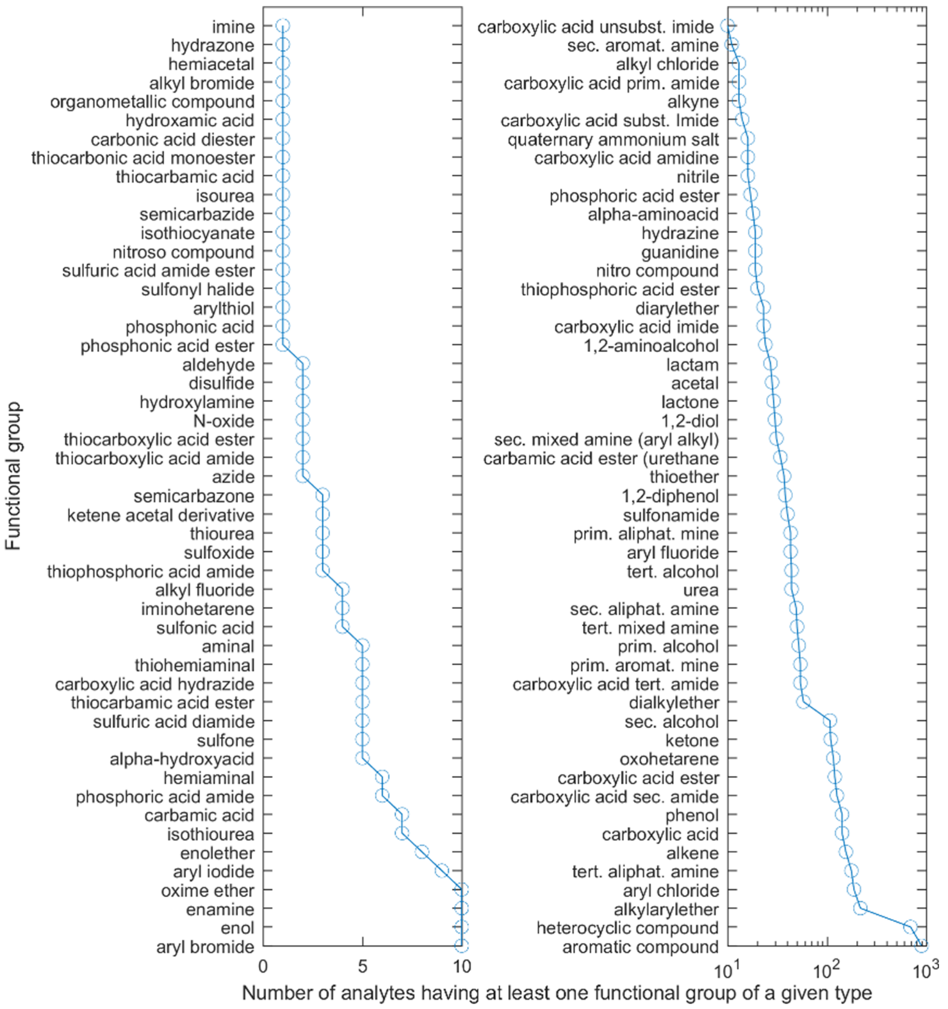


Figure 1. Functional groups identified by Checkmol. Figures show the number of analytes having at least one functional group of a given type.

vector regression, partial least squares (PLS), kernel-based PLS, least absolute shrinkage and selection operator, artificial neural networks, and random forests.[8]

In this study, we propose a Bayesian multilevel modeling framework as a tool that can build predictive chromatographic models. This idea is exemplified by proposing Bayesian multilevel modeling to describe relatively simple chromatographic data consisting of isocratic retention factor measurements. The general idea of this approach is as follows:

1 Characterize analytes using structural descriptors. We purposely focus on the number of various functional groups (substituents) and the molecular mass of analytes. These descriptors can be readily obtained from any analyte structure, practically without any cost. The effect of substituents can be easily interpreted in the sense that all individual analytes have different retentions, but any analyte would change retention in the same way if (counterfactually) it had a different molecular mass (due to the addition of certain hydrophobic fragments) or if it had one functional group replaced by another functional group.

2 Fit a Bayesian multilevel model. The multilevel model characterizes the entirety of the data using a single model. This model is based on (i) the same deterministic equation describing the relationship between the retention factor and organic modifier content for all considered analytes or more typically, any theoretically justified equation between the retention time and considered design variables; (ii) QSRRs relating the structure of the analyte (number of functional groups of a given type) and chromatography-specific parameters; and (iii) stochastic components of between-analyte, between-functional group, and residual (within-analyte) variability. Even in a large database of analytes, certain functional groups are rare. For such functional groups, it is difficult to estimate all regression coefficients precisely without any form of regularization. The multilevel models show an easy and intuitive way to implement hierarchical priors and consequently allow partial pooling of information across similar functional groups. This leads to less noisy estimates of the effects of functional groups on chromatographic parameters with sparse or even nonexistent data.[10]

3 Predict retention time/retention factor for analytes of interest. Predictions are straightforward as we are managing a model that generalizes to a heterogeneous group of analytes (i.e., with a wide range of functional groups). The Bayesian model allows us to propagate information shown experimentally and from prior assumptions using both existing theory and experience to the posterior predictions for any quantity of interest (i.e., retention times expected for a given set of analytes

for various chromatographic conditions). The resulting posterior distribution can be conveniently visualized in the same format as the chromatogram, yielding an uncertain chromatogram. This uncertainty chromatogram allows analysts to quickly assess the likely range of expected retention times/retention factors for given isocratic and gradient chromatographic conditions given all available information about the problem. The concept of using both prior and experimental data closely mimics the usual method development process in which one learns from experience and confirms what has been learned through experiments.[11]

This paper is organized as follows. In the following section, the data and model are presented using the standard statistical notation. Then, we present the inference results based on the model. Later, the usefulness of the model in predictions and their visualization as uncertainty chromatograms is shown using different types of preliminary data. We close with discussion and conclusions.

■ EXPERIMENTAL SECTION

Data. In this study, we used a publicly available data set with RP HPLC retention factor measurements collected for 1026 analytes (www.retentionprediction.org/hplc/database/). Retention times were measured under isocratic conditions on an Eclipse Plus C_{18} (Agilent) stationary phase with 3.5 μm particles. Experiments were conducted using 0.100% formic acid in water and 0.100% formic acid in acetonitrile as a mobile phase. The column temperature was 35 °C. Data were collected by Boswell et al.[12,13] and were used to create a method to predict retention time by back-calculating the gradient. The raw data are presented in Figure S1, and the molecular weight of the analytes ranged from 73.09 to 656.8 g/mol.

The molecular structure of the analytes was available in the SMILE format and was converted to the MDL mol format using OpenBabel.[14] Then, the input molecules were analyzed for the presence of approximately 204 functional groups and structural elements using Checkmol (version 0.5b N. Haider, University of Vienna, 2003−2018).[15] Functional groups that were not present on any analyte and functional groups merging other simpler functional groups were excluded from the analysis. In total, 100 unique functional groups were considered during model building. These functional groups and their frequency of occurrence are characterized in Figure 1.

Model. A nonlinear relationship between the decimal logarithm of the retention factor (log k) and organic modifier (Neue et al.[16] equation) was assumed to hold for all analytes

$$\log k_{ij} = \log k_{w,i} - \frac{S_{1,i}\cdot\varphi_j}{1 + S_{2,i}\cdot\varphi_j} \tag{1}$$

where log $k_{\mathrm{w},i}$, $S_{1,i}$, $S_{2,i}$ are the logarithm of the retention factor in water, the slope, and the curvature coefficient for the ith analyte, respectively, and φ_j denotes the jth acetonitrile content. For convenience, this equation was reparametrized to the retention factor in acetonitrile (log k_a) noticing that:

$$\log k_{a,i} = \log k_{w,i} - \frac{S_{1,i}}{1 + S_{2,i}} \tag{2}$$

The observed retention factors (log $k_{\mathrm{obs},z}$) were further modeled as

$$\log k_{\mathrm{obs},z} \sim \mathrm{student\_}t(\nu_{\mathrm{obs}},\ \log k_{i[z],j[z]},\ \sigma) \tag{3}$$

where z denotes the z^{th} measurement and Student_t denotes the Student's t-distribution with the mean given by eq 1, standard deviation σ, and normality parameter ν_{obs}. A tilde (~) denotes "has the probability distribution of" (i.e., the values of log $k_{\mathrm{obs},z}$ are randomly drawn from the given distribution—in this case, the Student's t-distribution). The Student's t-distribution was used to ensure robustness to outliers at the measurement level.

Multilevel modeling allows us to include a range of second-level models for analyte-specific parameters (log $k_{\mathrm{w},i}$, log $k_{\mathrm{a},i}$, and log $S_{2,i}$)

$$\begin{bmatrix} \log k_{w,i} \\ \log k_{a,i} \\ \log S_{2,i} \end{bmatrix} \sim \mathrm{MST}\left(\nu,\ \begin{matrix} \theta_{\log k_w + \beta_{\log k_w}\cdot(M_{\mathrm{mol},i} - 300)/100 - \pi_{\log k_w}\cdot X} \\ \theta_{\log k_a + \beta_{\log k_a}\cdot(M_{\mathrm{mol},i} - 300)/100 - \pi_{\log k_a}\cdot X} \\ \theta_{\log S_2 + \beta_{\log S_2}\cdot(M_{\mathrm{mol},i} - 300)/100 + \pi_{\log S_2}\cdot X} \end{matrix}\ ,\ \Omega\right) \tag{4}$$

where MST denotes the multivariate Student's t-distribution; θ is a vector of mean values of chromatographic parameters, where $\theta_{\log k_w}$, $\theta_{\log k_a}$, and $\theta_{\log S_2}$ denote the mean values of analyte-specific parameters for an analyte with the molecular mass of 300 and without any substituent, respectively; ν is a normality parameter; and Ω denotes a variance−covariance matrix. M_{mol} is the molecular mass, β is an effect of the molecular mass/100, where 100 is approximately the standard deviation of the available molecular masses of analytes, and π is an effect of each functional group on chromatographic parameters with separate values for log k_w, log k_a, and log S_2. In other words, π represents the difference in chromatographic parameters due to the presence of a functional group, assuming all else being equal. X is a matrix of size 1026 × 100 that decodes the number of functional groups present on each analyte. The lack of a particular functional group was denoted as 0, and the presence of a functional group was denoted as n, with n denoting the number of functional groups of the same type present on each analyte. S_2 was modeled on a logarithmic scale to ensure that S_2 values were positive.

Also, we decomposed the covariance matrix into a scale (ω) and a correlation (matrix ρ) based on the formula to ease the specification of the prior distribution

$$\Omega = \mathrm{diag}(\omega)\cdot\rho\cdot\mathrm{diag}(\omega) \tag{5}$$

Finally, a third-level model was used for regression parameters describing the effects of substituents ($\pi_{\log S_2}$, $\pi_{\log k_\mathrm{w}}$, and $\pi_{\mathrm{dlog}k}$ equal to the difference between $\pi_{\log k_\mathrm{w}}$ and $\pi_{\log k_\mathrm{a}}$)

$$\pi_{\log k_w,1:100} \sim \log\mathrm{normal}(\ln(\theta_{\pi\log k_w}),\ \sigma_{\pi\log k_w}) \tag{6}$$

$$\pi_{\mathrm{dlog}k,1:100} \sim \mathrm{student\_}t(\nu_\pi,\ \theta_{\pi\mathrm{dlog}k},\ \sigma_{\pi\mathrm{dlog}k}) \tag{7}$$

$$\pi_{\log S_2,1:100} \sim N(0,\ \sigma_{\pi\log S_2}) \tag{8}$$

where $\theta\pi$ denotes the effect of a typical functional group, and σ_π is a standard deviation of the individual $\pi_{1:100}$ values. In this study, $\pi_{\log k_\mathrm{w}}$ was restricted to be positive using a lognormal distribution. Basically, all functional groups identified by Checkmol are involved in ion, hydrogen bonding, dipole−dipole, dipole−induced dipole, and electron pair donor−electron pair acceptor interactions with the mobile-phase and stationary-phase constituents. Thus, analyte retention can be decreased in water-rich mobile phases, however, to different degrees. For convenience, the difference between the effects of functional groups in water-rich and acetonitrile-rich mobile phases $\pi_{\mathrm{dlog}k}$ ($\pi_{\log k_w}-\pi_{\log k_a}$) was modeled instead of $\pi_{\log k_a}$. The between-group variation of that quantity was assumed to follow the Student's *t*-distribution and assumes that the effect of a functional group in water and acetonitrile ($\pi_{\log k_w}$ and $\pi_{\log k_a}$) is correlated. The Student's *t*-distribution also ensures robustness because certain functional groups can have different retention characteristics in methanol than in acetonitrile. The symmetric distribution was selected for $\pi_{\log S_2}$.

Priors were formulated by calculating the approximate values of log $k_{w,i}$ and log $k_{a,i}$ using the least-square procedure and using the Neue model (eq 2) with S_2 = 2. This assumption was necessary to obtain stable estimates for all analytes. The values of log $k_{w,i}$ and log $k_{a,i}$ were then correlated with the molecular mass. This approach will be referred henceforth as a two-stage approach. The intercepts that correspond to the analyte with a molecular mass of 300 equal 3.6 and −1.7 and slopes equal to 1.4 and 0.2 for log k_w and log k_a, respectively. The standard deviation of the unexplained (i.e., residual) variability for log $k_{w,i}$ and log $k_{a,i}$ equals approximately 1.5. $\theta_{\log k_w}$ and $\theta_{\log k_a}$ were assumed to be two standard deviations higher than these calculated means, indicating that functional groups decrease analyte retention. In the case of $\theta_{\log S_2}$ parameters, the priors' mean of log(2) was based on data in the literature.[2,16,17] Also, we assumed a standard deviation of 0.2 on a logarithmic scale with a base of 10, corresponds to an a priori range of S_2 values from 0.9 to 4.3 (5th–95th percentile). The scale for $\beta_{\log k_w}$ and $\beta_{\log k_a}$ was 1.5 and that for $\beta_{\log S_2}$ was 0.2

$$\theta_{\log k_w} \sim N(6.6,\ 1.5) \tag{9}$$

$$\theta_{\log k_a} \sim N(1.3,\ 1.5) \tag{10}$$

$$\theta_{\log S_{2a}} \sim N(\log(2),\ 0.2) \tag{11}$$

$$\beta_{\log k_w} \sim N(1.4,\ 1.5) \tag{12}$$

$$\beta_{\log k_a} \sim N(0.2,\ 1.5) \tag{13}$$

$$\beta_{\log S_2} \sim N(0,\ 0.2) \tag{14}$$

Priors for residual variability (σ and ν_{obs}) equal

$$\sigma \sim N_{+}(0,\ 0.067) \tag{15}$$

$$\nu_{\mathrm{obs}} \sim \mathrm{gamma}(2,\ 0.1) \tag{16}$$

A scale of 0.067 was obtained from the residuals observed during the two-stage approach. The parameters for a gamma distribution for a normality parameter were selected to favor a normal distribution. The parameters ω, ρ, and ν were given the following priors:

$$\omega_{\log k_w} \sim N_{+}(0,\ 1.50) \tag{17}$$

$$\omega_{\log k_a} \sim N_{+}(0,\ 1.50) \tag{18}$$

$$\omega_{\log S_2} \sim N_{+}(0,\ 0.2) \tag{19}$$

$$\rho \sim \mathrm{LKJ}(3)(3 \times 3\ \mathrm{matrix}) \tag{20}$$

$$\upsilon \sim \mathrm{gamma}(2,\ 0.1) \tag{21}$$

where N_{+} denotes the half-normal distribution and LKJ denotes the Lewandowski et al.[18] distributions. In this case, LKJ(3) ensures that the density is uniform over correlation matrices of order 3.

The hyperparameters for the location and scale describing between-functional group variations were given simple hyperpriors, assuming derived scales of 1.5 for log k_w and log k_a and 0.2 for log S_2

$$\theta_{\pi-\log k_w} \sim N_{+}(0,\ 1.5),\ \theta_{\pi-\mathrm{dlog}k} \sim N(0,\ 1.5) \tag{22}$$

$$\sigma_{\pi-\log k_w},\ \sigma_{\pi-\mathrm{dlog}k} \sim N_{+}(0,\ 1.5),\ \sigma_{\pi-\log S_2} \sim N_{+}(0,\ 0.2) \tag{23}$$

$$\nu_{\pi} \sim \mathrm{gamma}(2,\ 0.1) \tag{24}$$

Simulations. Population-level parameters can be used in predictions as these parameters and the underlying deterministic part of the model store common analyte and functional group information about retention. This information can also be combined with any number of experiments to yield individualized chromatographic predictions for any analyte to which the model is expected to generalize. Posterior predictive checks were performed to assess the accuracy of such predictions. Predictive checks are simply a replicated data set using the model. These replicated data sets, when compared visually with the original data, allow us to assess model fit and the predictive capabilities of the model.[19] Additionally, a 10-fold leave-analyte-out cross-validation was used to assess model performance for a new analyte with a selected number of preliminary data. The analytes from the original data were randomly partitioned into 10 subsamples. Of the 10 subsamples, a single subsample was excluded from the analysis. The remaining nine subsamples plus zero or a limited number of measurements from the excluded analytes were used to obtain predictions for those excluded analytes. The cross-validation process was then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The isocratic retention factors were calculated using eq 1. The gradient retention times were calculated using the solution to the general gradient equation

$$\int_0^{t_R-t_0} \frac{\mathrm{d}t}{t_0 k_i(\phi(t))} = 1 \tag{25}$$

where $k_i(t)$ is the instantaneous retention factor corresponding to the isocratic retention factor, *k*, which would be obtained with the mobile phase composition actually present at a column inlet (eq 1), and t_0 is a column hold-up time. The predictions were obtained for a 20 min linear acetonitrile gradient with acetonitrile content changing from 0 to 1. A dwell time of 0.2 min was assumed for simulations. This equation was solved numerically using the trapezoidal method.

The posterior predictions were summarized as an uncertainty chromatogram. This chromatogram summarizes the posterior distribution as a probability density estimate of retention times/retention factors expected for a given set of analytes chromatographed under given conditions (and conditional on the available model and data). Such a chromatogram visualizes the uncertainty for the location of each peak maximum on a given chromatogram. Such a "peak" has a convenient interpretation: among all similar analytes (with respect to the molecular mass and functional groups, the area under the probability density function represents the

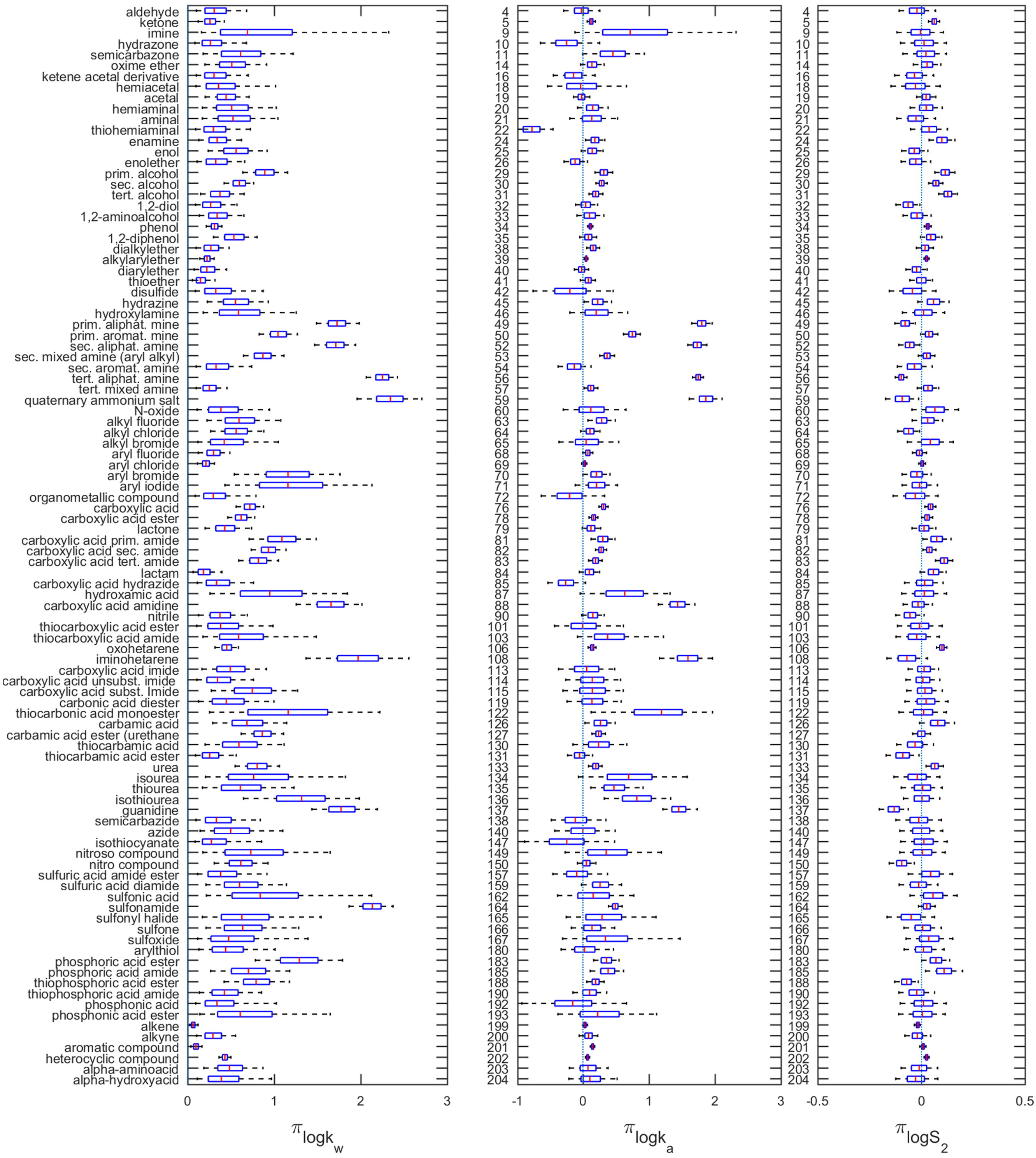


Figure 2. Graphical display of the marginal posterior distributions for the effects of each functional group on log k_w, log k_a, and log $S_{2.}$

fraction of analytes that are expected to have a retention factor/retention time within the range that area was calculated. Uncertainty decreases whenever there is access to any additional piece of information about the analyte (i.e., additional measurement).

Implementation. Multilevel modeling was performed in Stan/CmdStan 2.18 software[20] linked with MATLAB R2017b (The MathWorks, Inc., Natick, Massachusetts, United States) using MATLABStan 2.15 (Stan Development Team. 2017. MATLABStan: the MATLAB interface to Stan, http://mc-stan.org). For the application and simulation calculations, we used the following values of the Stan parameters: number of iterations = 1000, warm up = 1000, and number of Markov chains = 4. Stan codes were inspired by the work of Margossian and Gillespie.[21] Convergence diagnostics were checked using Gelman−Rubin statistics and trace plots. No divergence is reported in the model. The MATLAB code, data,

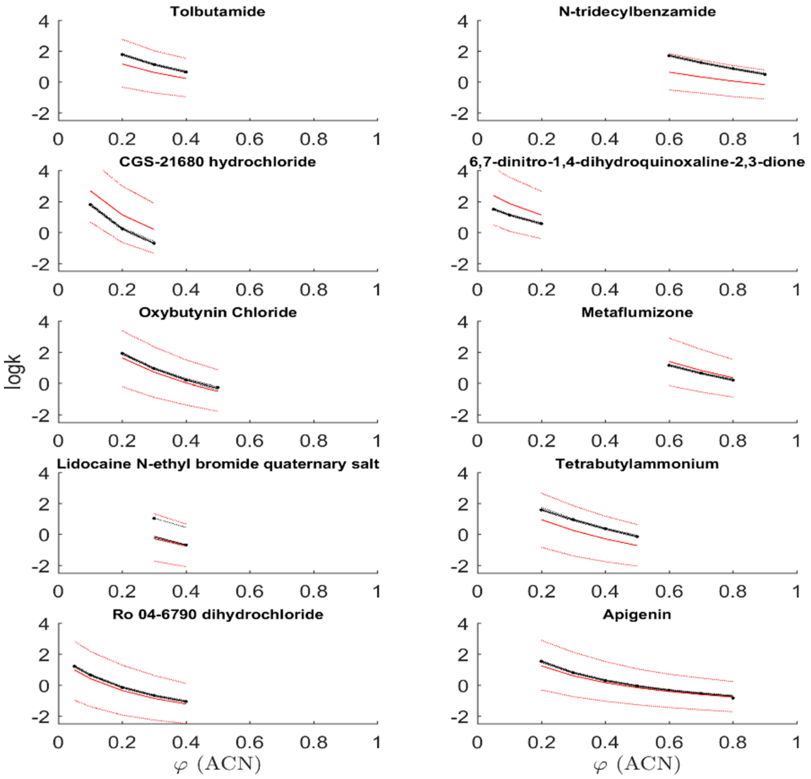


Figure 3. Individual and population predictions represented as posterior medians (lines) and 5th−95th percentiles (dotted lines) for a random set of 10 analytes. Observed retention factors are shown as dots. Black corresponds to future observations on the same analyte, and red corresponds to future observations of a new analyte.

and Stan code used to analyze the data are publicly available on GitHub (http://github.com/wiczling/bmm).

■ RESULTS

Multilevel modeling is a generalization of regression modeling in which model parameters are also given probability models.[10] This implies that model parameters are allowed to vary by group (i.e., by an analyte or by a functional group). In this study, this characteristic is exemplified by assuming the between-analyte variability of chromatographic parameters and between-functional group variability of regression coefficients. Consequently, the model comprises several nested models: (i) a measurement model, (ii) a model for analyte-specific chromatographic parameters, and (iii) a model for functional group effects (regression coefficients for functional group effects).

Table S1 shows a summary of the marginal posterior distributions for all population-level parameters. The effects of each functional group on a particular chromatographic parameter are visualized in Figure 2. The ordered values are also shown in Figure S2A−D to identify functional groups with the highest and lowest effects. The distribution of analyte-specific chromatographic parameters is given in Figure S3 (either directly or as eta plots, which show unexplained variability). Eta denotes the difference between the individual and expected values for a particular chromatographic parameter.

The typical values of $\theta_{\log k_w}$, $\theta_{\log k_a}$, and $\theta_{\log S_2}$ for an analyte without any functional group and molecular mass of 300 equal 7.10 (CI: 6.80−7.40), 0.05 (CI: −0.07−0.17), and 0.38 (CI: 0.34−0.38) on average, respectively. Log k_w is higher for analytes with a higher molecular mass by approximately 2.60 (CI: 2.50−2.80) per 100 g/mol difference in the molecular mass. The effect is less prevalent for log k_a. Specifically, log k_a is higher for analytes with higher molecular mass by approximately 0.40 (CI: 0.35−0.45) per 100 g/mol difference in the molecular mass. The effect of molecular mass on $\theta_{\log S_2}$ is negligible [−0.02 (CI: −0.04−0.00)]. Figure S4 visualizes the effect of molecular mass on analyte-specific chromatographic parameters. The standard deviation of the unexplained variability by functional groups and molecular mass amounts to 0.73 (CI: 0.67−0.78), 0.33 (CI: 0.30−0.35), and 0.12 (CI: 01−0.13) for log k_w, log k_a, and log S_2, respectively, with a small normality parameter of 2.30 (CI: 2.00−2.60) that indicates the presence of outliers (a nonnormal distribution). The correlation between analyte-specific chromatographic parameters was the highest between log k_w and log k_a (0.62, CI: 0.57−0.67). One of the important features of the multilevel model is the ability to regularize parameters (i.e., the analyte-specific chromatographic parameters) and is visible by comparing the parameters obtained using the two-stage approach and using the multilevel model. In the multilevel model, the individual values "shrink" toward typical values, as shown in Figure S5 because individual estimates are less "noisy" because they are constrained by the higher-level portion of the model.

In the water-rich mobile phase, functional groups decreased retention with a median decrease of 0.50 (CI: 0.40−0.61) per functional group and a coefficient of variation of approximately 100% (≈sqrt(exp(0.8$\hat{3}$2) − 1)*100). This high coefficient of variation indicates that the effects of functional groups on the retention factor in water-rich mobile phases vary considerably. As expected, functional groups represent a wide range of interactions, such as ion, hydrogen bonding, dipole−dipole,

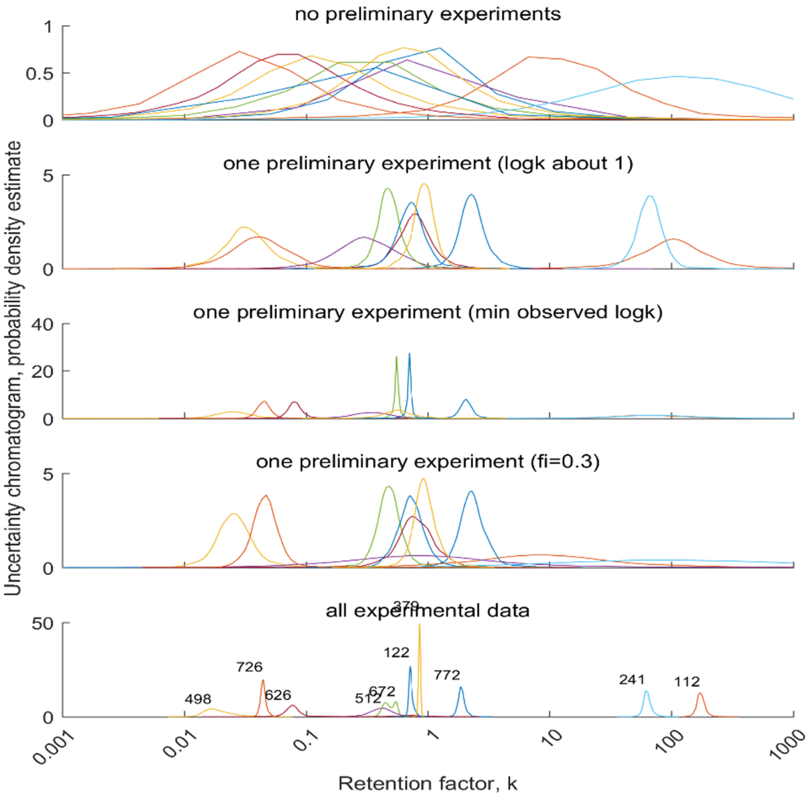


Figure 4. Uncertainty chromatograms summarizing predictions for a selection of isocratic conditions. Each peak represents the range of analyte retention factors along with uncertainty, as predicted by the proposed model conditional on different preliminary data. Colors correspond to different analytes that are identified at the bottom figure: 112: *N*-tridecylbenzamide, 122: tetrabutylammonium, 241: metaflumizone, 379: apigenin, 498: CGS-21680 hydrochloride, 512: 6,7-dinitro-1,4-dihydroquinoxaline-2,3-dione, 626: lidocaine *N*-ethyl bromide quaternary salt, 672: oxybutynin chloride, 726: Ro 04−6790 dihydrochloride, 772: tolbutamide.

dipol−induced dipol, and electron pair donor−electron pair acceptor interactions. The effect of a typical functional group is lower in acetonitrile than in the water-rich mobile phases by approximately 0.38 (CI: 0.30−0.47). The standard deviation for between-functional group variations for that difference was approximately 0.25 (CI: 0.17−0.34). The effects of functional groups that were not present can be approximated from these population-level parameters. It is of clear importance because it allows the model to generalize to other analytes with functional groups for which the experimental data are lacking. The between functional group variations for $\pi_{\log k_w}$, $\pi_{\mathrm{dlog}k}$, and $\pi_{\log S_2}$ are visualized in Figure S6.

Based on Checkmol notation, quaternary ammonium salt, tertiary aliphatic amine, and sulfonamide functional groups are among the substituents with the highest effects (>2) on log k_w values (these are likely ionized). As expected, the effect is the lowest for alkene and aromatic structures. The quaternary ammonium salt, aliphatic amines (primary, secondary, and tertiary), iminohetarene, and guanidine functional groups are among the substituents with the highest effects (>1) for log k_a and are also likely ionized in acetonitrile-rich mobile phases. The lowest effect was observed for thioheminal and carboxylic acid hydrazine. The effects of functional groups on log S_2 are fairly small. Several functional groups (e.g., sulfonamide) exhibit considerably different retentions in water-rich than in acetonitrile-rich mobile phases. Clearly, the effects of functional groups that are not common in the analyzed data set (imine, hydrazone, and hemiacetal) are predicted with large uncertainty. These high uncertainty intervals indicate a lack of knowledge about these parameters beyond that arising from the similarity of these functional groups to other functional groups.

Figure 3 shows the individual and population predictions for 10 randomly selected analytes. Individual predictions correspond to the future observations of the same analyte, and population predictions correspond to future observations of a new analyte. As expected, the individual predictions are highly accurate because they are based on population-level parameters, the number of functional groups, molecular mass, and all observed log k measurements for these analytes.

This characteristic is not true when predicting retention factors for an analyte for which no experimental data are available. Such typical predictions are uncertain because there is less information about the retention factor (only population-level parameters, number of functional groups, and molecular mass). The appropriate goodness-of-plots are presented to show the calibration and sharpness of the model predictions in Figure S7. These plots also summarize the accuracy of predictions expected after cross-validation, specifically individual predictions approximate leave-one-measurement-out cross-validation, and population predictions correspond to leave-one-analyte-out cross-validation; this occurs because the population-level parameters are typically insensitive to the lack of a single observation (individual predictions) or all observations for a particular analyte (population predictions).

It is critical to develop models that can predict analyte retention given access to different sizes of experimental data that can correctly propagate uncertainty for any quantity of interest (i.e., retention under gradient or isocratic conditions).

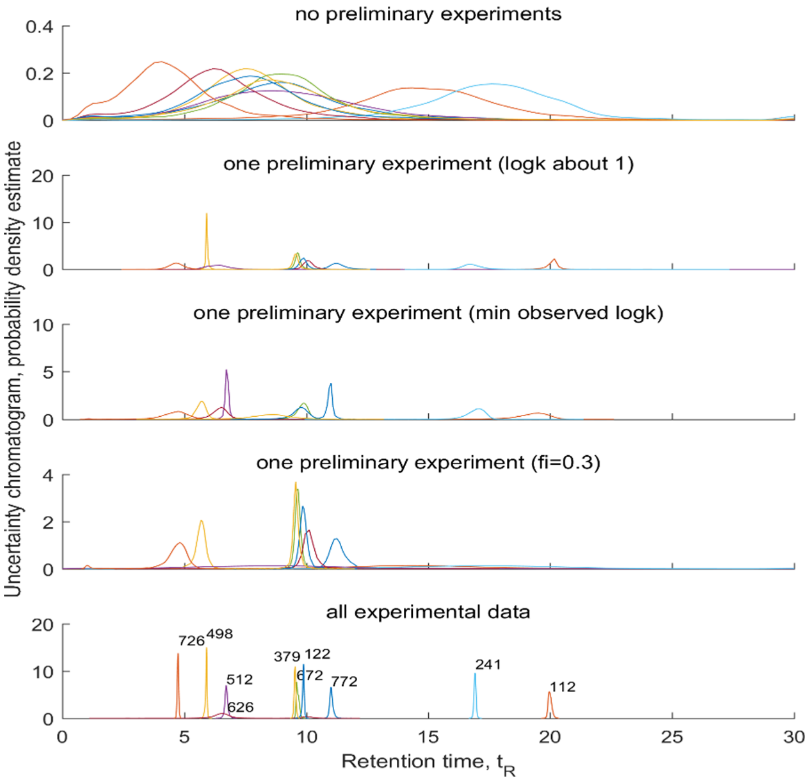


Figure 5. Uncertainty chromatograms summarizing prediction for a selection of gradient conditions. Each peak represents the range of analyte retention factors along with uncertainty, as predicted by the proposed model conditional on different preliminary data. Colors correspond to different analytes that are the same as in Figure 4.

The advantage of using multilevel models in this regard is clear. The model allows us to predict retention factors/retention time for a new analyte that is not present in the data set, the same analyte under the new chromatographic condition given excess to all the experimental data, or for an analyte with a different functional group and is also able to predict such behavior for a group of analytes belonging to a certain class, such as sulfonamides. The benefits of the multilevel model in predictions were shown for a random set of analytes, given the different preliminary data (no preliminary experimental, isocratic experiment with log k approximately 1, isocratic measurement with minimum observed value of log k for a particular analyte, one isocratic measurement conducted at φ = 0.3, and all analyte-specific experimental data) under isocratic and gradient conditions. The predictions for five scenarios are shown in Figures 4 (for isocratic) and 5 (for gradient conditions) for a random set of analytes and visualized as an uncertainty chromatogram. The predictions are also visualized using a different graphical display in Figure S8. At the beginning of any method development process when no experimental data are available, predictions can be expected to be uncertain. Figures 4 and 5 show that the information shown by functional groups and population-level parameters is of limited practical usefulness due to large uncertainty about that prediction. However, the uncertainty is finite, suggesting that certain information was obtained and that certain conclusions with regard to whether the separation of analytes is possible can be made even without any preliminary data. For example, if certain peaks on an uncertainity chromatogram do not overlap, it is unlikely for those compounds to have similar retention factor/retention times. If they do overlap, the chance of similar retention times increases. The situation changes when certain experimental data are added to the predictions. Even a single experiment can typically add a lot of information and thereby reduce uncertainty. This reduction in uncertainty is clearly different for each compound, type of preliminary experiment, and chromatographic condition. As can be expected, access to all analyte-specific observations leads to reasonably precise predictions. Outside of the considered situation, the model allows us to obtain well calibrated posteriors in the sense that if we repeat the process for a new analyte, the retention time/retention factor will fall in a 50% posterior interval, exactly 50% of the time.[22]

■ DISCUSSION

The proposed multilevel model was built to recapture the rules with which the data have been generated (mechanistic model) and allows for a direct and easy interpretation of all model parameters. The proposed model also shows a unified description of the entire data set. This characteristic is converse to classical methods of analyzing chromatographic data (e.g., the two-stage approach), which tend to ignore the hierarchical structure of the data and perform analysis at the analyte level. The Bayesian model requires defining the prior distribution for all model parameters and is an important part of the model that allows us to incorporate previous knowledge about the studied phenomenon and/or regularize inferences. In this study, an effort was made to define fairly weakly informative priors for all model parameters. Because priors are an important model assumption, they can be criticized and changed depending on different states of knowledge about the problem. For example, functional groups can be divided into subgroups (ionized, not ionized) with separate priors.

An analogous problem to that of finding a desired separation is encountered in the field of population pharmacokinetics.[23] Before obtaining any concentration measurement for a given patient, the only way to predict the pharmacokinetic profile and select the appropriate dose for that patient is to compare the patient to other similar patients (with similar covariates). When experimental data are available, this additional knowledge can be incorporated into the prediction of the patient-specific concentration profile and dose. Because there is some analogy between dose finding and searching for the desired separation in chromatography, similar tools might be used. When managing limited preliminary data in chromatography, it is necessary to compare the analytes in the sample to other similar analytes that were analyzed before. Similarity might be described by various descriptors, such as the number of functional groups. Each time new experimental data are available, the model should be sufficiently flexible to incorporate this additional piece of information into the predictions and decision making. The multilevel model presented in this study also shares some analogy with the multilevel regression and poststratification approach used in political studies.[24] The Bayesian methods also become popular in assisting the end user to make decisions with respect to the presence/absence of the xenobiotics in the LC−mass spectrometry spectrum.[25]

The multilevel model naturally groups analytes into different types based on the number of functional groups; thus, a large number of localized QSRRs are built for each combination of functional groups. The ability of the multilevel model to pool information allowed us to obtain stable estimates of model parameters that would otherwise be difficult to achieve without large data sets. It is a different approach from that proposed by Haddad et al.[26] under the name of a localized quantitative QSRR modeling approach. Their algorithm finds analytes from certain (possibly large) databases that are similar to analytes for which the predictions are desired. The similarity of analytes is then defined through a given similarity measure, such as the structural similarity of compounds (i.e., Tanimoto similarity index), the physicochemical similarity of compounds (i.e., lipophilicity), the neutral, acidic, or the basic nature of the compound. Then, a QSRR model is built based on this restricted data set that is further used to predict retention for the desired group of analytes. In this case, localized predictions might be noisy, particularly if there are only a few similar analytes in the data set available for predictions. Clearly, a direct comparison of both approaches is required to fully assess method performance under different scenarios.

The chemical space is large; thus, building large databases of retention times is difficult.[27] For most chromatographic experiments, analytes are selected based on convenience and availability; thus, special attention must be paid to the generalizability of model predictions. In this study, the proposed model should generalize well with other analytes if their molecular mass is lower than 600 g/mol, regardless of the number and type of functional groups. In this analysis, a large data set of analytes was considered; however, there were still functional groups that were not present on any analyte. Despite this limitation, the proposed model allows effective management of "rear" functional groups. Simply, the effects of such substituents (π values) were estimated with large uncertainty. Conversely, the effects of a functional group for groups that were common in the analyzed data were estimated more precisely. Additionally, the large amount of isocratic data conducted at one pH value does not allow us to fully elucidate the effects of pH on retention. This effect certainly occurs for certain analytes in this data set.[28,29] Isocratic or gradient experiments conducted at different pH values of the mobile phases are necessary to fully account for this additional complexity. Regardless of the situation, the model can predict chromatographic parameters and retention of analytes and may serve as a template to solve more complex problems encountered in chromatography and related fields. Such an approach allows for the "individualized" prediction of retention time, as shown in Figures 4 and 5.

In this study, the effect of each functional group on chromatographic parameters can be understood as a Hansch constant (the hydrophobic parameter for a specific substituent).[30] The effects of substituents on the parameters describe the interaction of small analytes with various macromolecules and chromatographic stationary phases. This concept has a long tradition. The basic idea is reflected by the following formula: $\log P_{\mathrm{R-X}} = \log P_{\mathrm{R-H}} + \pi_{\mathrm{R}}$, where R−X is an analyte, log P is lipophilicity, and R is a substituent. The π_{R} once known can be useful to extrapolate the known log P of a parent analyte to that with a substituent. This equation works if the log P of the parent analyte (R−H) is known, and there is not much interaction with other groups present in the parent compound.[31] In this study, the parent analyte describes an analyte without any functional group, and it is assumed that analytes without any functional group have different retentions that depend on the analyte molecular mass. In this setting, the effect of a functional group has a counterfactual interpretation because replacing one group with another will lead to changes in retention equal to the difference in π values. The use of a multilevel model for estimating the Hansch constant has the advantage of quantifying the uncertainty for each constant. The same reasoning as presented in this study for chromatographic retention could be extended to other problems involving relationships between an analyte structure and activity.

In our opinion, it is neither impossible to use QSRR nor to use QSRR with confidence to find sufficiently precise conditions, leading to the desired separation in situations of limited experimental data. The QSRR equations explain certain part of between-analyte variability; however, the accuracy of such prediction is rather limited when only population-level parameters (i.e., QSRR equations) are used for predictions. However, even in this case, the retention time/retention factor can be predicted, and knowledge about the likely chromatogram can be deduced. Information about the analyte structure is helpful and should be incorporated into decision making if possible.[3,4] The expected retention times, given the various sources of information (population-level parameters, functional groups, molecular mass, and any measurements), can be easily visualized, giving analysts a tool to make rapid decisions with regard to further steps: (i) stop method development if the desired separation is improbable, (ii) do more experimentation if current information is uncertain, or (iii) claim that this method is sufficient (i.e., performing more experiments is unlikely to show a better separation). Uncertainty chromatograms can also be used to quantify the probability of successful separation[5] or to calculate the expected utility[4] of the next experiment.

Multilevel models are a new concept in the field of chromatography. However, recent computational advances allow for easy implementation of these methods in practice.

https://doi.org/10.1021/acs.analchem.0c05227

Many complexities and improvements can be made to make this model more general and useful in daily practice [e.g., peak width, other descriptors, a range of columns, more diverse chromatographic conditions (including methanol, acetonitrile, pH, and temperature), or between-laboratory variability]. To achieve this, more collaboration with regard to the collection of chromatographic data is definitely required. Bayesian multilevel modeling appears to be a tool that provides a unified scheme for analyzing large chromatographic databases.

■ CONCLUSIONS

A Bayesian multilevel framework was used to build a chromatographic model describing isocratic retention factor measurements for 1026 heterogeneous analytes. This modeling approach allowed us (i) to naturally account for the hierarchical structure of the data, (ii) to estimate the effect of many functional groups on chromatographic parameters, and (iii) to predict retention for new (not yet analyzed) analytes or analytes for which only a limited number of measurements were available. The Bayesian-based multilevel model also allowed us to incorporate prior knowledge and use known chromatographic theory in predictions. Under this framework, it was possible to calculate the uncertainty of predictions and visualize them as an uncertainty chromatogram (e.g., posterior probability density of retention factors expected for each analyte under given chromatographic conditions, given the current knowledge about the retention). Such a visualization of predictions was discussed in terms of usefulness in decision making. We are aware that the model is complex but it is built out of simple, easily understandable blocks. A modern state-of-the-art platform for statistical modeling and a high-performance statistical computation environment are available, making this approach an interesting alternative to various chemometric procedures.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c05227.

> Summary of the MCMC simulations of the marginal posterior distributions of population-level model parameters; relationship between the logarithm of the retention factor (log k) and acetonitrile content in the mobile phase; graphical display of the marginal posterior distributions for the effects of each functional group on $\pi_{\log k_w}$, $\pi_{\log k_a}$, $\pi_{\log S_2}$, and $\pi_{\log k_w}-\pi_{\log k_a}$; scatter plots between individual chromatographic parameters or eta values (difference between the analyte-specific chromatographic parameter and expected value) and molecular mass; effect of molecular mass on retention of compounds without functional groups; comparison of model parameters obtained using the two-stage approach and multilevel model; histogram of mean posterior values of the effects of each functional group on chromatographic parameters; goodness-of-fit plots; and predictions for a random set of 10 analytes (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Paweł Wiczling − *Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, 80-416 Gdańsk, Poland*; orcid.org/0000-0002-2878-3161; Email: wiczling@gumed.edu.pl

Authors

Agnieszka Kamedulska − *Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, 80-416 Gdańsk, Poland*

Łukasz Kubik − *Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, 80-416 Gdańsk, Poland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c05227

Author Contributions

Ł.K. and A.K. prepared the data, P.W. designed the study, P.W. and A.K. analyzed the data, and P.W., Ł.K., and A.K. wrote the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was supported by the National Science Centre, Poland (grant 2015/18/E/ST4/00449).

■ DEDICATION

This article is dedicated to the memory of prof. Roman Kaliszan (1945−2019).

■ REFERENCES

(1) Kaliszan, R. *Chem. Rev.* **2007**, *107*, 3212−3246.

(2) Snyder, L.; Kirkland, J.; Dolan, J. *Introduction to Modern Liquid Chromatography*; Wiley: Hoboken, N.J., 2010.

(3) Kubik, Ł.; Kaliszan, R.; Wiczling, P. *Anal. Chem.* **2018**, *90*, 13670−13679.

(4) Wiczling, P. *Sep. Sci. plus* **2018**, *1*, 63−75.

(5) Wiczling, P.; Kaliszan, R. *Anal. Chem.* **2016**, *88*, 997−1002.

(6) Briskot, T.; Stückler, F.; Wittkopp, F.; Williams, C.; Yang, J.; Konrad, S.; Doninger, K.; Griesbach, J.; Bennecke, M.; Hepbildikler, S.; Hubbuch, J. *J. Chromatogr. A* **2019**, *1587*, 101−110.

(7) Héberger, K. *J. Chromatogr. A* **2007**, *1158*, 273−305.

(8) Haddad, P. R.; Taraji, M.; Szücs, R.Prediction of Analyte Retention Time in Liquid Chromatography. *Anal. Chem.* **2020**, *93*. DOI: 10.1021/acs.analchem.0c04190.

(9) Žuvela, P.; Skoczylas, M.; Jay Liu, J.; Bączek, T.; Kaliszan, R.; Wong, M. W.; Buszewski, B. *Chem. Rev.* **2019**, *119*, 3674−3729.

(10) Gelman, A. *Technometrics* **2006**, *48*, 432−435.

(11) Box, G. E. P. *J. Appl. Stat.* **1996**, *23*, 3−20.

(12) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr. A* **2011**, *1218*, 6742−6749.

(13) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr. A* **2011**, *1218*, 6732−6741.

(14) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 33.

(15) Haider, N. *Molecules* **2010**, *15*, 5079−5092.

(16) Neue, U. D.; Phoebe, C. H.; Tran, K.; Cheng, Y.-F.; Lu, Z. *J. Chromatogr. A* **2001**, *925*, 49−67.

(17) Pappa-Louisi, A.; Nikitas, P.; Balkatzopoulou, P.; Malliakas, C. *J. Chromatogr. A* **2004**, *1033*, 29−41.

(18) Lewandowski, D.; Kurowicka, D.; Joe, H. *J. Multivariate Anal.* **2009**, *100*, 1989−2001.

(19) Gelman, A.; Carlin, J. B.; Stern, H. S.; Rubin, D. B. *Bayesian Data Analysis*; Chapman & Hall/CRC: Boca Raton, 2004.

(20) Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M.; Guo, J.; Li, P.; Riddell, A. *J. Stat. Software* **2017**, *76*, 1−32.

(21) Margossian, C.; Gillespie, B. Differential Equations Based Models in Stan. http://mc-stan.org/events/stancon2017-notebooks/stancon2017-margossian-gillespie-ode.html. (accessed November 19, 2000).

(22) Cook, S. R.; Gelman, A.; Rubin, D. B. *J. Comput. Graph Stat.* **2006**, *15*, 675−692.

(23) Mould, D. R.; Upton, R. N. *CPT Pharmacometrics Syst. Pharmacol.* **2012**, *1*, No. e6.

(24) Hanretty, C. *Polit. Stud. Rev.* **2020**, *18*, 630−645.

(25) Woldegebriel, M.; Vivó-Truyols, G. *Anal. Chem.* **2015**, *87*, 7345−7355.

(26) Wen, Y.; Talebi, M.; Amos, R. I. J.; Szucs, R.; Dolan, J. W.; Pohl, C. A.; Haddad, P. R. *J. Chromatogr. A* **2018**, *1541*, 1−11.

(27) Haddad, P. *LC GC* **2017**, *35*, 499−502.

(28) Wiczling, P. *Anal. Bioanal. Chem.* **2018**, *410*, 3905−3915.

(29) Kaliszan, R.; Wiczling, P. *TrAC, Trends Anal. Chem.* **2011**, *30*, 1372−1381.

(30) Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991**, *91*, 165−195.

(31) Leo, A.; Jow, P. Y. C.; Silipo, C.; Hansch, C. *J. Med. Chem.* **1975**, *18*, 865−868.