Gdański Uniwersytet Medyczny
Wydział Farmaceutyczny

mgr Agnieszka Kamedulska

Opracowanie bayesowskich modeli hierarchicznych opisujących retencję w wysokosprawnej chromatografii cieczowej w odwróconym układzie faz

Development of Bayesian hierarchical models describing retention in reversed-phase high-performance liquid chromatography

Praca wykonana
w Zakładzie Biofarmacji i Farmakokinetyki
Katedry Biofarmacji i Farmakodynamiki
Gdańskiego Uniwersytetu Medycznego

Promotor pracy:
prof. dr hab. Paweł Wiczling

Gdańsk, 2023

Pragnę serdecznie podziękować mojemu promotorowi
Prof. dr. hab. Pawłowi Wiczlingowi za całą wiedzę
przekazaną mi przez te lata oraz wszelką pomoc.
Dziękuję za niezwykle cenne wskazówki,
poświęcony czas i wyrozumiałość.

Osobne podziękowania składam na ręce mojego męża Bartosza,
który wspierał mnie na każdym etapie pracy nad rozprawą.
Słowa podziękowania należą się także moim Rodzicom,
na których zawsze mogłam liczyć.

ŹRÓDŁA FINANSOWANIA

Badania stanowiące przedmiot niniejszej rozprawy doktorskiej były finansowane ze środków przyznanych w ramach:

- projektu NCN SONATA BIS 5 (2015/18/E/ST4/00449) "Optymalizacja warunków rozdzieleń chromatograficznych z wykorzystaniem nieliniowego modelowania efektów mieszanych i technik bayesowskich",
- projektu POWR.03.02.00-00-I035/16-00 współfinansowanego przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Wiedza Edukacja Rozwój 2014-2020.

Obliczenia prowadzone na potrzeby rozprawy doktorskiej wykonano z wykorzystaniem komputerów Centrum Informatycznego Trójmiejskiej Akademickiej Sieci Komputerowej w ramach grantu obliczeniowego.

Spis treści

Streszczenie

Wysokosprawna chromatografia cieczowa w odwróconym układzie faz (RP-HPLC) jest najczęściej wykorzystywaną techniką chromatograficzną. Jest to związane z jej zdolnością do separacji wieloskładnikowych mieszanin związków chemicznych o podobnych właściwościach. Ważnym elementem poprzedzającym analizę chromatograficzną jest konieczność ustalenia warunków prowadzenia rozdzielenia. Do tego celu można wykorzystywać modele matematyczne, których parametry estymuje się w oparciu o dane pochodzące z serii eksperymentów wstępnych. Jednak nadal rutynowym podejściem w poszukiwaniu optymalnych warunków rozdzieleń jest tzw. metoda prób i błędów.

Analiza wieloskładnikowych mieszanin za pomocą chromatografii cieczowej sprzężonej ze spektrometrią mas (LC/MS) pozwala uzyskać duże zbiory danych retencyjnych. Dane te są stosunkowo łatwe do zebrania, jednak zwykle cechują się heterogenicznością, dużą liczbą wartości odstających i mogą zawierać brakujące rekordy lub nawet błędy systematyczne, powodowane brakiem precyzyjnej kontroli nad eksperymentami, instrumentami i procesem zbierania danych. Analiza danych musi być poprzedzona czyszczeniem i filtrowaniem. Dodatkowo do uzyskania przydatnych informacji wymagane są stosunkowo złożone modele, np. ze względu na obecność analitów o różnej charakterystyce retencji i występowanie różnych źródeł zmienności. Do opisu takich zbiorów danych można wykorzystać modele empiryczne (statystyczne) lub modele mechanistyczne (generatywne). Modele empiryczne są zwykle budowane w oparciu jedynie o wiedzę statystyczną bez powiązania z dostępną teorią chromatograficzną. Z drugiej strony modele mechanistyczne opisują proces generujący dane, który bazuje na teorii chromatografii cieczowej. Takie modele są bardziej odpowiednie do ekstrapolacji. Do przewidywań retencji najczęściej wykorzystuje się modele regresji, których parametry oszacowuje się algorytmami uczenia maszynowego. Niezależnie od podejścia, dokładne przewidywanie czasu retencji w chromatografii cieczowej jest wymagane do szybkiej przesiewowej oceny kolumn, wspomaganego komputerowo opracowywania i przenoszenia metod oraz jednoznacznej identyfikacji związków za pomocą analiz LC/MS.

Celem niniejszej rozprawy doktorskiej była analiza danych chromatograficznych na podstawie bayesowskich modeli hierarchicznych. Zastosowana metodyka pozwoliła uwzględnić wiedzę aprioryczną dostępną w literaturze oraz scharakteryzować niepewności parametrów modelu i różnego rodzaju predykcji.

Pierwszy model miał na celu charakterystykę retencji analitów przy użyciu deskryptorów strukturalnych takich jak rodzaj grupy funkcyjnej i masa molowa. Model opracowano w oparciu o dane zebrane dla 1026 analitów w warunkach izokratycznych (dla różnych zawartości acetonitrylu w fazie ruchomej). Podczas opracowywania modelu powstał koncept chromatogramu niepewności. Na tym chromatogramie każdy pik reprezentuje retencję analitu wraz z niepewnością przewidywaną przez proponowany model w zależności od różnych danych wstępnych.

Drugi model został opracowany na podstawie tych samych danych. Celem tej

analizy była ocena użyteczności $\log P$ i pK_a jako predyktorów. Wykazano, że informacje dostarczane przez $\log P$ i pK_a są niewystarczające do precyzyjnego przewidywania współczynnika retencji. Jednak dodanie informacji z jednego eksperymentu wstępnego pozwala na obniżenie tej niepewności i poprawę precyzji predykcji.

W kolejnym modelu wykorzystano dane obejmujące 187 analitów mierzonych w warunkach gradientowych zarówno w acetonitrylu jak i w metanolu przy różnych wartościach pH fazy ruchomej, temperatury i czasów trwania gradientu. Przedstawiony model pozwolił na charakterystykę retencji analitów obojętnych, kwasowych i zasadowych dla szerokiego zakresu warunków chromatograficznych. Model ten stanowił również przyczynek do problemu porównania kolumn oraz podejmowania decyzji związanych z wyznaczeniem warunków chromatograficznych prowadzących do uzyskania żądanego rozdzielenia.

Zaproponowane w pracy modele są interpretowalne i zapewniają zwięzłe podsumowanie złożonych danych. Mogą służyć do przewidywania niepewności retencji na podstawie różnej liczby wstępnych eksperymentów oraz być przydatne do podejmowania decyzji w warunkach niepewności. Mogą również dostarczyć informacji wstępnych dla kolejnych analiz. Tym samym, bayesowskie modele hierarchiczne wydaję się być interesującą alternatywą dla różnych procedur chemometrycznych i metod uczenia maszynowego wykorzystywanych w analizie danych chromatograficznych.

Słowa kluczowe: modelowanie retencji, model wielopoziomowy, wnioskowanie bayesowskie, rozwój metody

Abstract

Reversed-phase high-performance liquid chromatography (RP-HPLC) is a popular analytical technique due to its ability to separate multi-component mixtures of analytes. It also provides large datasets of retention times, especially when coupled with mass spectrometry detection (LC/MS). These data are relatively easy to collect, but tend to be heterogeneous, messy, and may contain missing records or even systematic errors due to a lack of precise control over experiments, instruments, and the data collection process. Analysis of such data sets often requires pre-processing in the form of cleaning and filtering. In addition, relatively complex models are required to extract useful information, e.g. due to the presence of analytes with different retention characteristics and the presence of different sources of variation. Empirical (statistical) or mechanistic (generative) models are often used to describe such datasets. Empirical models are built based on statistical knowledge without connection to the available chromatographic theory (e.g. using machine learning algorithms). Mechanistic models, on the other hand, are derived from the principles and foundations of liquid chromatography. Such models are more suitable for extrapolation. Regardless of the approach, accurate retention time prediction in liquid chromatography is required for rapid column screening, computer-aided method development and portability, and unequivocal identification of compounds by LC/MS analyses.

This doctoral dissertation aimed to describe the retention of analytes in RP-HPLC using Bayesian hierarchical models. The applied methodology allows the inclusion of measures of prior knowledge in the analyses and quantification of the uncertainty of model parameters and predictions.

The first model aimed to characterize the retention of analytes using structural descriptors such as type and number of functional groups and molecular mass. The model was developed based on data collected for 1026 analytes in isocratic conditions (for different acetonitrile contents in the mobile phase). During the model development, the concept of the uncertainty chromatogram was introduced. In this chromatogram, each peak represents the analyte retention along with the uncertainty predicted by the proposed model depending on various input data.

The second model was developed based on the same dataset. The purpose of this analysis was to assess the usefulness of $\log P$ and pK_a as predictors. It was shown that the information provided by $\log P$ and pK_a has limited added predictive value. However, adding information from one preliminary experiment reduces this uncertainty and improves the accuracy of predictions.

The third model was developed based on data from 187 analytes measured in gradient conditions in both acetonitrile and methanol at different pH values of the mobile phase, temperature, and for a range of gradient durations. This model allowed for the characterization of the retention of neutral, acidic, and basic analytes for a wide range of chromatographic conditions. This model can be also used to illustrate the

application of the hierarchical models for column comparison and in the search for chromatographic conditions leading to the desired separation.

The developed models are interpretable and provide a concise summary of complex data. They can be used to predict retention uncertainty based on a varying number of preliminary experiments and be useful for decision-making under uncertainty. They can also provide *a priori* information for subsequent analyses. Thus, Bayesian hierarchical models seem to be an interesting alternative to various chemometric and machine-learning methods used in the analysis of chromatographic data.

Keywords: retention modeling, multilevel model, Bayesian inference, method development

Lista publikacji

Niniejsza rozprawa doktorska oparta jest na cyklu artykułów, który stanowią trzy prace oryginalne. W jego skład wchodzą następujące pozycje:

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

B. Kamedulska, A., Kubik, Ł., Wiczling, P. (2022). Statistical analysis of isocratic chromatographic data using Bayesian modeling. Analytical and Bioanalytical Chemistry, 414(11), 3471–3481. https://doi.org/10.1007/s00216-022-03968-x

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

Pozostałe publikacje niebędące częścią rozprawy doktorskiej:

- Popowicz, H., Mędrzycka-Dąbrowska, W., Kwiecień-Jaguś, K., Kamedulska, A. (2021). Knowledge and Practices in Neonatal Pain Management of Nurses Employed in Hospitals with Different Levels of Referral—Multicenter Study. Healthcare, 9(1), 48. https://doi.org/10.3390/healthcare9010048
- Ryś, D., Jaczewski, M., Pszczoła, M., Kamedulska, A., Kamedulski, B. (2022). Factors affecting low-temperature cracking of asphalt pavements: analysis of field observations using the ordered logistic model. International Journal of Pavement Engineering, 1–11. https://doi.org/10.1080/10298436.2022.2065273

Rozdział 1
Wstęp

Obecnie wysokosprawna chromatografia cieczowa w odwróconym układzie faz (RP-HPLC) jest najpopularniejszą techniką chromatograficzną. Szacuje się, że jest wykorzystywana w ponad 90% wszystkich separacji [Žuvela et al. (2019)]. Znajduje ona szerokie zastosowanie w wielu obszarach nauki, m.in. w farmacji (badania leków), w badaniach środowiskowych (biomonitoring zanieczyszczeń), w kryminologii (ilościowe oznaczanie narkotyków w próbkach biologicznych), w badaniach klinicznych (ilościowe oznaczanie stężeń leków) czy w branży spożywczej (analizy konserwantów, pomiarów jakości wody itp.).

Rozdzielenie jest możliwe dzięki różnemu czasowi przebywania poszczególnych analitów w kolumnie chromatograficznej. Podczas eksperymentu RP-HPLC pompa wysokociśnieniowa, która stanowi system dostarczania rozpuszczalnika, generuje określone szybkości przepływu fazy ruchomej. Dozownik wprowadza próbkę do stale płynącego strumienia fazy ruchomej, która przenosi próbkę do kolumny HPLC. Kolumna ta wypełniona jest fazą stacjonarną. Różna dystrybucja analitów pomiędzy fazą stacjonarną i ruchomą pozwala na oddzielenie pasm związków eluujących z kolumny HPLC. Do identyfikacji tych pasm niezbędny jest detektor, którym zwykle jest spektrometr mas.

Ważnymi parametrami charakteryzującymi proces chromatograficzny są retencja, selektywność i rozdzielczość. Retencją określa się zjawisko wolniejszej niż szybkość przepływu eluentu migracji poszczególnych analitów z próbki i wyraża się poprzez współczynnik retencji. Gdy każdy składnik rozdzielanej mieszaniny wykazuje zróżnicowaną retencję, to układ chromatograficzny nazywa się selektywnym wobec analitów. Selektywność kolumny chromatograficznej opisywana jest poprzez stosunek współczynników retencji sąsiadujących ze sobą pików. Z kolei rozdzielczość określa stopień separacji pików.

W przypadku RP-HPLC faza ruchoma jest bardziej polarna od fazy stacjonarnej i stanowi ją najczęściej woda/bufor z dodatkiem modyfikatora organicznego: acetonitrylu (ACN), metanolu (MeOH) lub tetrahydrofuranu (THF). Z kolei faza stacjonarna składa się zwykle z immobilizowanych na krzemionce różnej długości łańcuchów węglowodorowych, w szczególności C_{18} i C_8 [Snyder et al. (1997)].

Szczegółowe mechanizmy retencji warunkujące separację nie są do końca poznane [Nikitas et al. (2002)]. Badania prowadzone na poziomie molekularnym, które pozwoliłyby na ich pełne zrozumienie, są obarczone dużymi trudnościami. Wynika to m.in. z tego, że dane dotyczące retencji są pomiarami termodynamicznymi, zatem nie mogą one dać molekularnego obrazu retencji. Mogą jedynie służyć do wnioskowania o typowym zachowaniu się analitów w układzie chromatograficznym. Z kolei badania spektroskopowe, mogące dać bardziej szczegółowy wgląd, nie pozwalają zaobserwować dokładnych konfiguracji pojedynczej cząsteczki analitu, a cały złożony rozkład możliwych konfi-

guracji [Rafferty et al. (2007)]. Jednak zastosowanie narzędzi takich jak symulacje Monte Carlo (MC) lub dynamika molekularna (MD) wydaje się być obiecujące do szczegółowego wyjaśnienia i interpretacji mechanizmów determinujących retencję na poziomie molekularnym [Rafferty et al. (2007); Gritti (2021)].

Wiadomo jednak, że mechanizm retencji w RP-HPLC opiera się jednocześnie na dwóch procesach, a mianowicie adsorpcji i podziale [Gritti i Guiochon (2005a,b)]. Na potrzeby ich rozróżnienia adsorpcja oznacza, że analit jest w kontakcie powierzchniowym z fazą stacjonarną, natomiast podział, że analit jest w przybliżeniu całkowicie osadzony w fazie stacjonarnej [Dorsey i Dill (1989)].

Własności RP-HPLC sprawiają, że może być ona stosowana zarówno w badaniach jakościowych, jak i ilościowych, to znaczy służyć do identyfikacji oraz oznaczania ilościowego szerokiej gamy związków chemicznych. Prawie wszystkie rozdzielenia HPLC można przeprowadzić w odwróconym układzie faz w porównaniu z mniej uniwersalną HPLC w normalnym układzie faz (NP-HPLC). Istnieje tylko kilka typów rozdzieleń, w których RP-HPLC jest mało skuteczna. Należą do nich separacja jonów nieorganicznych (do której wykorzystuje się chromatografię jonowymienną), polisacharydów (będących zbyt hydrofilowymi związkami, aby mogła zachodzić adsorpcja w fazie stałej), polinukleotydów (adsorbujących się nieodwracalnie na powierzchni fazy stacjonarnej), a także związków silnie hydrofobowych (ze względu na dużą retencję). Jednak oprócz wspomnianych wyjątków ma ona szerokie zastosowanie, począwszy od węglowodorów prostych i aromatycznych, po proste aminy, cukry, lipidy, czy związki farmakologicznie czynne. Jest również stosowana do rozdzielania aminokwasów, peptydów i białek, a także cząsteczek pochodzenia biologicznego [Skoog et al. (2017)].

W dalszej części pracy zostaną przedstawione podstawowe składowe i własności RP-HPLC, począwszy od omówienia próbki i jej znaczenia w kontekście separacji, przedstawienia wpływu poszczególnych faz na retencję oraz istotności doboru trybu elucji, po wskazanie metod wykorzystywanych do predykcji retencji.

1.1 Próbka

Wiedza o składzie chemicznym próbki może dostarczyć cennych wskazówek ułatwiających dobór warunków rozdzielenia. Jednymi z ważniejszych informacji o próbce są liczba zawartych w niej związków, ich rozpuszczalność, struktura chemiczna, masa molowa, wartość pK_a oraz zakres stężeń [Snyder et al. (1997)].

Liczba analitów w analizowanej próbce ma kluczowe znaczenie w szczególności, gdy konieczne jest rozdzielenie wszystkich związków chemicznych. Zwykle pełna separacja to bardzo trudny do przeprowadzenia proces, podczas gdy rozdzielenie mniejszego podzbioru analitów jest często dużo prostsze. W przypadku dużej liczby związków analizowanych w jednym eksperymencie istotnym czynnikiem staje się czas. Często pożądane jest zmniejszenie czasu przebiegu kosztem spadku rozdzielczości [Snyder et al. (1997)].

Kolejnym istotnym czynnikiem wpływającym na retencję w systemie RP-HPLC jest struktura chemiczna związków zawartych w próbce. Zwykle wartości czasu retencji t_R zmieniają się w regularny i przewidywalny sposób wraz z wielokrotnym podstawieniem pewnej grupy w cząsteczce próbki (jak np. w szeregu homologów czy benzologów). Wprowadzenie danej grupy funkcyjnej do dowolnej cząsteczki próbki w danym układzie chromatograficznym zmieni jej retencję, modyfikując opisujący ją współczynnik retencji k o pewien stały współczynnik π_i. Te stałe są różne w zależności od podstawnika i na

ogół zależą od jego polarności. Wartości π_i maleją wraz z polaryzacją w systemach z odwróconymi fazami [Snyder et al. (2009)]. Na przykład niepolarne grupy funkcyjne, takie jak łańcuchy alkilowe ($-CH_3$ czy $-C_2H_5$) będą zwiększały czas retencji w niepolarnej fazie stacjonarnej. Natomiast bardziej polarne grupy funkcyjne, takie jak kwasy karboksylowe ($-COOH$) lub grupy aminowe ($-NH_2$), będą prowadziły do zmniejszenia czasu retencji.

Możliwe jest również skorelowanie kolejności elucji w chromatografii cieczowej w odwróconym układzie faz ze współczynnikiem podziału oktanol/woda $\log P$. $\log P$ jest miarą hydrofobowości związku i opisuje stosunek stężeń analitu (podział) pomiędzy dwoma niemieszającymi się fazami, fazą niepolarną (n-octanol) i fazą wodną (woda). Zwykle związki o wyższych wartościach $\log P$, wskazujących na wyższą hydrofobowość, będą miały dłuższe czasy retencji na kolumnie RP-HPLC, ponieważ będą silniej oddziaływać z niepolarną fazą stacjonarną. Jednak należy zauważyć, że związek między $\log P$ a retencją w RP-HPLC nie zawsze jest prosty i może na niego wpływać wiele czynników [Hanai (1991)].

Masa cząsteczkowa jest również jednym z czynników, które mogą wpływać na retencję w układzie RP-HPLC. Jednak jej wpływ jest generalnie mniej znaczący w porównaniu z innymi czynnikami, takimi jak hydrofobowość związku i jego interakcje z fazą stacjonarną. Na ogół mniejsze cząsteczki będą miały krótsze czasy retencji niż większe cząsteczki, ponieważ mają mniejszą powierzchnię do interakcji z fazą stacjonarną. Jednak efekt ten nie zawsze jest liniowy [Snyder et al. (2009)].

RP-HPLC jest zwykle przeprowadzana w warunkach kwasowych. W takich warunkach związki zasadowe będą zjonizowane i mają krótsze czasy retencji, ponieważ w mniejszym stopniu oddziałują z hydrofobową fazą stacjonarną. Natomiast związki kwasowe będą przeważnie niezjonizowane i będą miały dłuższe czasy retencji na kolumnie RP-HPLC ze względu na ich zwiększoną hydrofobowość i oddziaływanie z fazą stacjonarną [Dong (2006)].

Stężenie związków w badanej próbce powinno być dobrane pod kątem pojemności fazy stacjonarnej. Ta natomiast zależy m.in. od jej pola powierzchni i hydrofobowości oraz odnosi się do ilości substancji rozpuszczonej, która może zostać zatrzymana przez kolumnę przed nasyceniem. Przy niskich stężeniach pojemność fazy stacjonarnej nie jest przekraczana, a czasy retencji związków są względnie stałe. Jednak przy wysokich stężeniach pojemność fazy stacjonarnej może zostać przekroczona, a czasy retencji związków mogą się zmniejszyć [Snyder et al. (2009)].

Rozpuszczalność analitów w próbce może również znacząco oddziaływać na retencję. Wynika to z jej wpływu na stężenie analitów w fazie ruchomej. Jeśli anality nie są w pełni rozpuszczalne w fazie ruchomej, mogą tworzyć cząstki lub wytrącać się, co może prowadzić do zmniejszenia stężenia i uniemożliwić ich detekcję [Snyder et al. (2009)]. Również oddziaływania analit-faza stacjonarna są podobne do tych, które odpowiadają za rozpuszczalność w danym rozpuszczalniku [Hanai (1999)].

1.2 Faza ruchoma

Fazę ruchomą stanowi rozpuszczalnik, który przesuwa zawarte w próbce anality przez kolumnę chromatograficzną. Faza ruchoma wchodzi w interakcje zarówno z fazą stacjonarną, jak i związkami w niej rozpuszczonymi, wskutek czego ma silny wpływ na retencję. Idealna faza ruchoma powinna charakteryzować się wysoką rozpuszczalnością dla próbki, niskim kosztem, wysoką czystością, przeźroczystością (w przypadku detektora

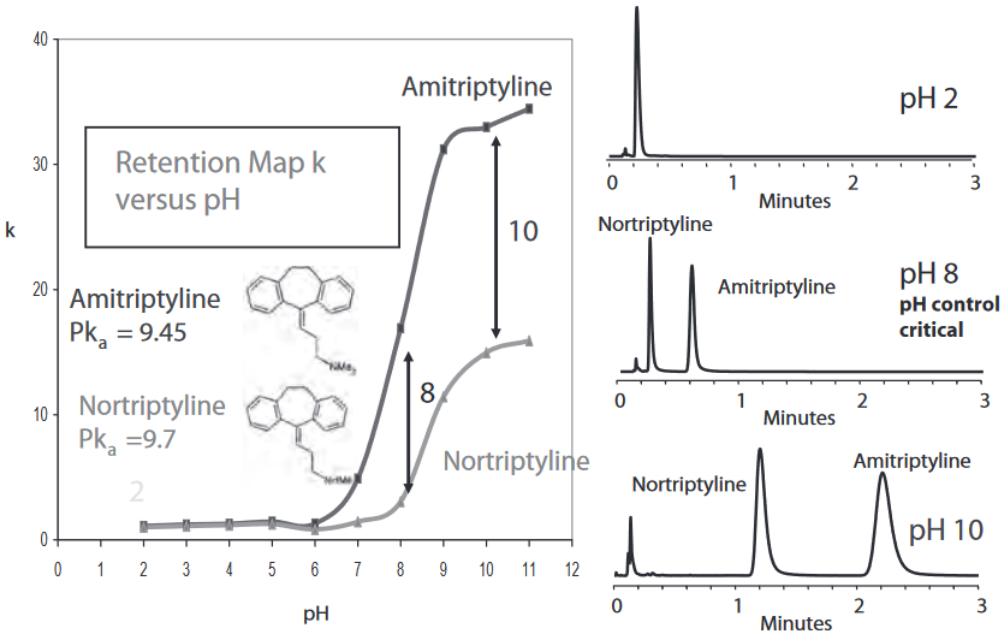
UV), a także nie powinna korodować elementów aparatury HPLC. Pożądanymi cechami są również niska lepkość, mała toksyczność oraz niepalność.

Charakterystyczną własnością każdego rozpuszczalnika jest jego siła elucyjna. Określa ona zdolność rozpuszczalnika do wymywania analitów z kolumny i wiąże się z jego polarnością. W przypadku RP-HPLC, z uwagi na hydrofobowość fazy stacjonarnej, woda jest rozpuszczalnikiem o małej sile elucyjnej, w przeciwieństwie do rozpuszczalników organicznych, których siła elucyjna maleje w kolejności:

$$\text{THF} > \text{ACN} > \text{MeOH} \gg \text{woda}.$$

Mała moc rozpuszczalnika, jakim jest woda, wynika ze słabej rozpuszczalności związków organicznych [Dong (2006)].

Wartość pH fazy ruchomej ma kluczowy wpływ na retencję związków jonizowalnych (kwasów i zasad). W chromatografii cieczowej w odwróconym układzie faz zjonizowana forma analitu ma znacznie niższy współczynnik retencji niż postać obojętna. Zostało to zilustrowane na Rysunku 1 na przykładzie dwóch strukturalnie podobnych analitów (amitryptyliny i nortryptyliny). Możemy zauważyć, że przy pH=2 anality są zjoni-



Rysunek 1: Wykres po lewej przedstawia zależność współczynnika retencji k od pH fazy ruchomej przy stałej zawartości procentowej modyfikatora organicznego dla amitryptyliny i nortryptyliny. Natomiast wykresy po prawej przedstawiają chromatogramy uzyskane w trzech różnych pH.
Źródło: Dong M. W., Modern HPLC for Practicing Scientists, Wiley-Interscience 2006, rozdz. 2, s. 31.

zowane i eluują jako pojedynczy pik. Przy pH=8 anality są częściowo zjonizowane i dobrze rozdzielone. Z kolei przy pH=10 obie te substancje są niezjonizowane i silnie zatrzymywane, dzięki czemu możliwe jest ich rozdzielenie. Stąd niezwykle istotna jest kontrola pH w układzie chromatograficznym. Do tego celu wykorzystuje się bufory. Do opracowywania metod HPLC sprzężonych ze spektrometrem mas (MS) używa się m.in. buforów soli amonowych lotnych kwasów. Jednak należy pamiętać, że bufory są skuteczne tylko w zakresie $\pm 1,5$ jednostki pH od ich pK_a.

W RP-HPLC często stosuje się pH w zakresie 2,5-3. Takie pH hamuje jonizację analitów będących słabymi kwasami, dzięki czemu ich retencja jest większa. Kolejną zaletą niskiego pH jest brak jonizacji grup silanolowych, co redukuje osadzanie się substancji zasadowych na powierzchni fazy stacjonarnej. Jednak rozwój faz stacjonarnych obecnie pozwala na stosowanie również wysokich wartości pH (nawet na poziomie 12). Pozwala to na rozdzielenie dwóch blisko spokrewnionych leków, amitryptyliny i nortryptyliny, które prezentuje Rysunek 1. W niskim pH oba anality są zjonizowane i eluują wspólnie z czołem rozpuszczalnika. Natomiast w pH zbliżonym do pK_a analitów, są one częściowo zjonizowane i dzięki temu dobrze rozdzielane.

Kolejnymi ważnymi parametrami związanymi z fazą ruchomą są szybkość przepływu fazy ruchomej F i temperatura T. Wykorzystywanie wyższych szybkości przepływu F zwiększa przeciwciśnienie kolumny, jednocześnie skracając czas retencji i czas eksperymentu. Z kolei wyższe temperatury kolumny T obniżają lepkość fazy ruchomej, a co za tym idzie przeciwciśnienie kolumny, i zwykle znacząco wpływają na retencję, selektywność, czy sprawność kolumny. Ta ostatnia to własność kolumny umożliwiająca uzyskanie ,,ostrych'' pików i rozdzielanie wielu składników próbki w stosunkowo krótkim czasie [Dong (2006)].

1.3 Faza stacjonarna

Kolumna chromatograficzna jest ważnym elementem każdego systemu chromatograficznego. Podlega ona największym zmianom spośród wszystkich elementów układu chromatograficznego [Snyder et al. (2009)]. Zawiera ona drobne nośniki podtrzymujące fazę stacjonarną zapewniającą zróżnicowaną retencję poszczególnych składników próbki. Podstawowymi parametrami kolumny są jej tryb (RP-HPLC, NP-HPLC itd.), wymiary, typ (jak np. krzemionka czy polimer), cel zastosowania (chromatografia ilościowa lub jakościowa) oraz charakterystyka upakowania (wielkość cząstek i porów).

Praca ta skupia się na chromatografii cieczowej w odwróconym układzie faz, i to kolumnom tego trybu zostaną poświęcone dalsze rozważania. Wymiary kolumny tzn. jej długość i średnica pozwalają na kontrolę jej sprawności oraz na charakterystykę roboczą, tj. zakres możliwych do zastosowania szybkości przepływu czy wartości przeciwciśnienia. Dłuższe kolumny mają większą liczbę półek teoretycznych (większą sprawność), dzięki którym osiągają one lepszą rozdzielczość przy dłuższym czasie analizy. Również spadek ciśnienia w kolumnie jest proporcjonalny do jej długości. Jednak w przypadku prostych mieszanin próbek, krótsze kolumny mogą zapewnić wystarczającą rozdzielczość i krótszy czas analizy [Dong (2006)].

Natomiast jeśli chodzi o typ kolumny, to najczęściej wykorzystywanym materiałem nośnym jest krzemionka. Kolumny wypełnione niezwiązaną krzemionką są rzadko używane do celów analitycznych ze względu na silne właściwości adsorpcyjne. Retencja zmienia się wraz z naturą nośnika. Ogólnie, im dłuższy jego łańcuch czy większa hydrofobowość, tym większa jest retencja. Stąd retencja w kolumnie z wykorzystaniem łańcuchów węglowodorowych C_{18} jest zwykle większa niż kolumnie z wykorzystaniem łańcuchów węglowodorowych C_8. Generalnie retencja w RP-HPLC związków niepolarnych, niezjonizowanych wzrasta w zależności od wypełnienia kolumny według następującego wzoru:

$$\text{niezwiązana krzemionka} \ll \text{cyjanowe} < C_4 < \text{fenylowe} < C_8 \approx C_{18}.$$

Z kolei polimerowe materiały nośne są wykorzystywane głównie w bioseparacjach i wspomaganiu chromatografii jonowymiennej. W porównaniu z krzemionką ich głównymi zaletami są szeroki zakres pH (1-14) i brak aktywnych grup silanolowych. Siła i wydajność kolumn z wypełnieniem polimerowym poprawiły się w ostatnich latach, jednak nadal mają gorszą sprawność niż fazy stacjonarne na bazie krzemionki [Snyder et al. (1997)].

Kolejnym istotnym aspektem jest rozmiar cząsteczek wypełnienia. Kolumny wypełnione małymi cząstkami zapewniają znacznie mniejszą utratę sprawności przy dużych szybkościach przepływu. Jednakże, ponieważ przeciwciśnienie kolumny jest odwrotnie proporcjonalne do kwadratu rozmiaru cząsteczek, kolumny wypełnione cząstkami poniżej 3 μm mogą powodować przekraczanie limitu ciśnienia większości urządzeń HPLC. Należy zauważyć, że zmniejszenie rozmiaru cząstek przy zachowaniu stałej długości kolumny może zwiększyć jej sprawność poprzez wzrost wysokości piku.

W chromatografii ilościowej zwykle stosuje się większe kolumny, aby pomieścić większe objętości próbek i umożliwić większe szybkości przepływu fazy ruchomej. Z kolei kolumny wykorzystywane w chromatografii jakościowej są mniejsze i mają wyższą rozdzielczość, co pozwala na lepszą separację i wykrywanie niewielkich różnic między związkami [Snyder et al. (2009)].

Większość podłoży chromatograficznych jest porowata, dzięki czemu zapewniają większą powierzchnię i pozwalają zmaksymalizować oddziaływanie analitów z fazą stacjonarną. Na ogół nośniki o dużej powierzchni prowadzą do większej retencji analitów. Natomiast upakowania o małych porach są problematyczne w przypadku dużych biomolekuł, które mogą zostać splątane lub uwięzione w porach, prowadząc do wolniejszego przenoszenia masy i dodatkowego poszerzenia pasma chromatograficznego [Dong (2006)].

Producenci wciąż opracowują nowe fazy stacjonarne o coraz lepszych własnościach, a mianowicie o większej wytrzymałości mechanicznej, lepszej odtwarzalności, szybszej kinetyce przenoszenia masy, wyższej sprawności, szerszym zakresie stabilności pH, czy wyższej selektywności. W ostatniej dekadzie nastąpił szybki wzrost liczby nowych stałych nośników, np. cząstek polimeru, cząstek hybrydowych, cząstek krzemionki pokrytych złożoną szczepioną warstwą organiczną. Produkty te rzeczywiście lepiej wykonują swoje zadania, jednak dokładny mechanizm rozdzieleń w nich nie jest znany [Gritti i Guiochon (2005b)].

1.4 Tryb elucji

Ważnym krokiem w opracowaniu metody RP-HPLC jest wybór trybu elucji, tj. elucji izokratycznej lub elucji gradientowej. Odpowiednio dobrany tryb pozwala na uzyskanie żądanego rozdzielenia w akceptowalnym czasie analizy.

1.4.1 Elucja izokratyczna

W separacji w warunkach izokratycznych faza ruchoma nie zmienia swojego składu przez cały czas trwania eksperymentu. Retencja w tych warunkach zwykle obliczana jest według następujących wzorów:

$$t_R = t_0(1 + k) \quad \text{lub} \quad V_R = V_0(1 + k), \tag{1.1}$$

gdzie t_R oznacza czas retencji, V_R objętość retencji, t_0 czas martwy, V_0 objętość martwą, a k współczynnik retencji.

Czas retencji jest to czas przebywania analitu w kolumnie chromatograficznej, odpowiada on elucji maksimum piku. Natomiast objętość retencji stanowi objętość potrzebną do elucji środka pasma danej substancji [Snyder et al. (2009)].

Z kolei czas martwy kolumny definiuje się jako czas od momentu wprowadzenia do kolumny substancji wnikającej do wszystkich porów wypełnienia, lecz nieulegającej sorpcji, aż do chwili pojawienia się maksimum piku tego związku na wylocie z kolumny. Parametr ten nie jest związany z procesem retencji, a zależy jedynie od szybkości przepływu fazy ruchomej F oraz objętości martwej kolumny:

$$t_0 = \frac{V_0}{F}.$$

Objętość martwa zależy od właściwości fizycznych kolumny (tj. długości, średnicy, porowatości fazy stacjonarnej) i wyraża objętość fazy ruchomej wewnątrz kolumny chromatograficznej [Snyder et al. (1997)]. Wyznaczenie objętości martwej nie należy do prostych, gdyż jej wartość uzyskiwana eksperymentalnie obarczona jest sporą niepewnością. Uznaje się, że błąd w estymacji tego parametru waha się do około 20%. Najczęściej do wyznaczenia czasu martwego kolumny w układzie faz odwróconych wykorzystuje się substancje takie jak: uracyl lub stężony roztwór azotanu sodu [Snyder et al. (1997)], które uznaje się za związki niezatrzymywane w kolumnie.

Natomiast współczynnik retencji k wyraża podział analitu między fazę stacjonarną i ruchomą oraz określany jest wzorem:

$$k = \frac{n_s}{n_m} = \frac{c_s \cdot V_s}{c_m \cdot V_m},$$

gdzie indeksy dolne s i m odnoszą się odpowiednio do fazy stacjonarnej i fazy ruchomej, a n oznacza ilość moli analitu, c stężenie analitu, V objętość fazy [Jarosz (2006)]. Jednak w przypadku chromatografii cieczowej z wykorzystaniem porowatych materiałów wypełniających, zdefiniowanie, co stanowi fazę stacjonarną, a co fazę ruchomą, jest niemożliwe [Knox i Kaliszan (1985)]. Współczynnik retencji k zależy w znaczącym stopniu od zawartości modyfikatora organicznego φ, która ma duży wpływ na siłę elucji i selektywność. Dodatkową zaletą jest elastyczność i dokładność w implementacji zmian tego czynnika. Nie istnieje precyzyjny wzór, który pozwalałby przewidywać współczynnik retencji na podstawie zawartości modyfikatora organicznego [Snyder et al. (1989)]. Natomiast jest wiele modeli wyrażonych przez przybliżenia takiej zależności, pozwalających oszacować wartości k. Do powszechnie stosowanych równań należy:

$$\log k = \log k_w - S \cdot \varphi,$$

gdzie k_w oznacza współczynnik retencji w czystej wodzie (na ogół jest to wartość ekstrapolowana), a S stałą charakterystyczną dla danego analitu i warunków chromatograficznych; opisuje siłę elucyjną modyfikatora dla tego konkretnego związku [Jarosz (2006)]. Równanie nazywane jest modelem liniowej siły elucyjnej (linear solvent strength model, LSS). Na ogół jednak zależność pomiędzy logarytmem współczynnika retencji i zawartością modyfikatora organicznego jest nieliniowa [Shoenmakers et al. (1979)].

Zaletą elucji izokratycznej jest łatwiejsze przenoszenie metod między kolumnami, instrumentami czy laboratoriami w porównaniu z elucją gradientową. W tym trybie

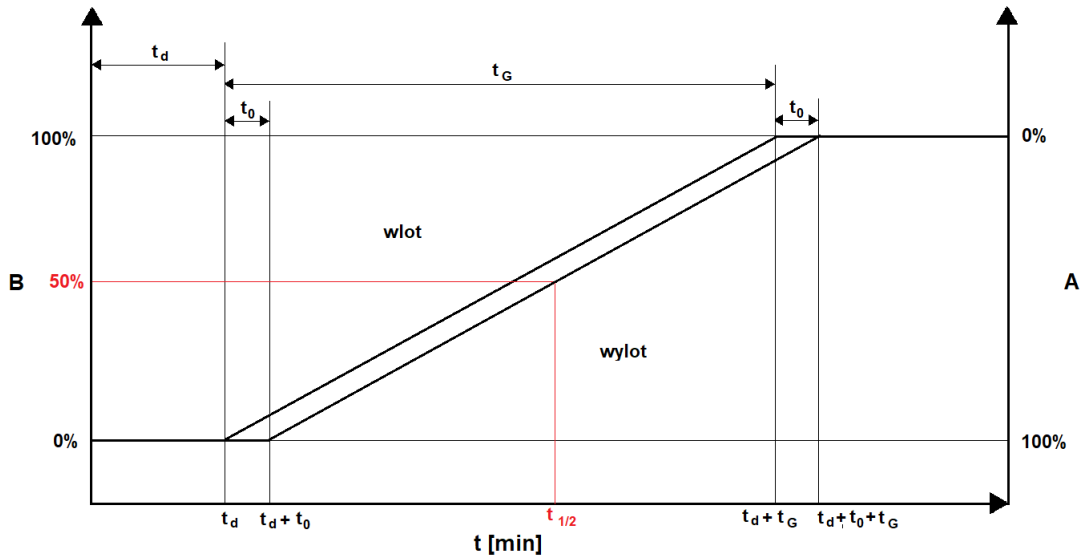
również prościej o optymalizację, gdyż mniej zmiennych wpływa na selektywność. Ponadto oprzyrządowanie jest mniej złożone i nie wymaga tak regularnej konserwacji jak w przypadku elucji gradientowej. Ogólnie uważa się, że elucja izokratyczna jest z natury szybsza, ponieważ nie wymaga przepłukiwania kolumny po każdym przebiegu w celu powrotu do początkowego składu fazy ruchomej. W przypadku elucji gradientowej istnieje wiele problemów, takich jak chociażby obecność pików niezwiązanych z przebiegiem trwającej analizy (tzw. ghost peaks) czy szum linii podstawowej. Mogą one utrudniać oznaczenia ilościowe i prowadzić do niedokładnych wartości powierzchni piku [Snyder et al. (1997); Schellinger i Carr (2006)].

Z drugiej strony wiele separacji nie jest możliwych przy wykorzystaniu elucji izokratycznej. Przyjmuje się, że współczynniki retencji badanej próbki nie mogą wykraczać poza przedział $(0.5, 15)$, a sama liczebność próbki nie powinna przekraczać 10 związków [Schellinger i Carr (2006)]. Charakteryzuje się ona również słabą rozdzielczością początkowych pików. A piki eluujące z opóźnieniem często są bardzo szerokie i płaskie [Snyder et al. (2009)]. To sprawia, że bardziej uniwersalna jest elucja gradientowa.

1.4.2 Elucja gradientowa

W elucji gradientowej skład fazy ruchomej zmienia się w trakcie trwania przebiegu chromatograficznego. Zwykle stosowane są dwa rozpuszczalniki A i B, z których rozpuszczalnik B mający większą siłę elucyjną zwiększa swoją zawartość w trakcie trwania przebiegu chromatograficznego. W wyniku tego retencja analitów mierzona przez współczynnik retencji k maleje [Snyder et al. (1997)].

Szybkość zmian w zawartości modyfikatora organicznego B w fazie ruchomej może być wyrażona przez różne krzywe. Jednak najczęściej wykorzystywany jest gradient liniowy, który pozwala na względnie prosty opis matematyczny retencji analitu. Rysunek 2 przedstawia przykładowe zmiany składu fazy ruchomej zarówno na wlocie, jak i wylocie kolumny chromatograficznej w trakcie trwania przebiegu. Rysunek ilustruje parametry



Rysunek 2: Przykładowy przebieg gradientu liniowego

związane z elucją gradientową, takie jak: t_d, który oznacza czas opóźnienia (*ang. dwell time*). Wielkość ta związana jest z unikalną dla instrumentu analitycznego objętością

układu od punktu, w którym mieszają się rozpuszczalniki fazy ruchomej, do początku kolumny. Jest to tak zwana objętość opóźnienia V_d (*ang. dwell volume*). Czas opóźnienia t_d jest ilorazem objętości opóźnienia V_d oraz szybkości przepływu fazy ruchomej F i wyraża czas niezbędny na dotarcie początku gradientu do wlotu kolumny. Mierzy się go m.in. poprzez wyznaczenie czasu do osiągnięcia 50% zawartości rozpuszczalnika B na wylocie kolumny, oznaczonego jako $t_{1/2}$, i odjęciu od niego połowy czasu trwania gradientu t_G:

$$t_d = t_{1/2} - \frac{1}{2}t_G.$$

Objętość opóźnienia ma praktyczne znaczenie jedynie w przypadku elucji gradientowej. Podczas separacji izokratycznych opóźnienie również występuje, jednak skład fazy ruchomej jest taki sam przez cały przebieg, stąd nie obserwuje się różnicy między chromatogramami z eksperymentów prowadzonych na instrumentach o różnych objętościach opóźnienia [Dolan (2006)].

W HPLC wyróżniamy jeszcze objętość pozakolumnową V_e (*ang. extra-column volume*). Definiuje się ją jako objętość między punktem dozowania próbki a punktem detekcji, z wyłączeniem kolumny. Składają się na nią objętości dozownika próbki, przewodów łączących i detektora [McNaught, A. D. i Wilkinson, A. (1997)]. Efekty pozakolumnowe mogą wpływać na szerokość piku, wydajność i rozdzielczość separacji. Jednak instrumenty posiadają opcje, które mogą redukować ich wpływ [Hong i McConville (2018)].

Kolejnym niewyjaśnionym wcześniej parametrem zobrazowanym na rysunku 2 jest czas trwania gradientu t_G. Określa on czas realizacji programu elucji przez instrument analityczny [Kamiński et al. (2004)] i stanowi podstawowy parametr opisujący gradient.

Współczynnik retencji w przypadku elucji gradientowej ma nieco inne znaczenie w porównaniu do techniki izokratycznej. W tej drugiej parametr k dla każdego związku jest stały przez cały czas trwania przebiegu. Natomiast w technice gradientowej wraz ze zmianą składu fazy ruchomej następuje zmiana jego wartości. Może być on wyrażony jako funkcja chwilowych izokratycznych wartości współczynników retencji w czasie trwania gradientu w zależności od składu fazy ruchomej w miejscu znajdowania się analitu [Jarosz (2006)] i oznaczony jako $k_i(t)$.

W przypadku gradientu liniowego czasy retencji analitów t_R możemy wyznaczyć poprzez zastosowanie podstawowego równania elucji gradientowej, które opiera się o model zakładający przejście nieskończenie małej objętości fazy ruchomej (o stałym składzie) dV przez środek pasma analitu w miarę migracji pasma przez kolumnę dx. Założenie to odpowiada wielostopniowemu gradientowi o nieskończonej liczbie stopni izokratycznych. W warunkach izokratycznych objętość retencji jest wyznaczana wzorem (1.1). Natomiast całkowita objętość fazy ruchomej przechodzącej przez środek pasma analitu nazywana jest jego zredukowaną objętością retencji V'_R i wyrażona wzorem: $V'_R = V_R - V_0 = V_0 \cdot k$. Stąd ułamkowy dystans pokonywany przez pasmo wynosi $dx = \frac{dV}{V_0 \cdot k}$. Gdy całkowita objętość fazy ruchomej V'_R przemieści się przez środek pasma ($\Sigma\, dV = V'_R$), to suma cząstkowych migracji wynosi 1 ($\Sigma\, dx = 1$). Stąd zachodzi:

$$\int_0^{V'_R - V_e} \frac{dV}{V_0 \cdot k(\varphi(t))} = 1. \tag{1.2}$$

Równanie (1.2) może być przeformułowane na terminy związane z czasem:

$$\int_0^{t'_R - t_e} \frac{dt}{t_0 \cdot k_i(t)} = 1$$

i wyraża ułamkową migrację pasma dx jako przyrost czasu dt podzielony przez tak zwany zredukowany czas retencji $t'_R = t_R - t_0$ [Snyder i Dolan (2006)].

W gradiencie liniowym φ zmienia się liniowo wraz z czasem t. W ogólności przyjmuje się:

$$\varphi = \begin{cases} \varphi_0 & \text{dla } t \leqslant t_d \\ \varphi_0 + \alpha \cdot t & \text{dla } t_d < t \leqslant t_g + t_d \\ \varphi_f & \text{dla } t > t_g + t_d \end{cases},$$

gdzie φ_0 odpowiada początkowej izokratycznej zawartości modyfikatora organicznego, α — nachyleniu gradientu, φ_f — końcowej zawartości modyfikatora organicznego [Nikitas i Pappa-Louisi (2005)].

1.5 Modele matematyczne

Przewidywanie retencji odgrywa ważną rolę w usprawnianiu procesu analitycznego, np. w wyznaczaniu warunków pozwalających uzyskać żądaną separację [Gritti (2021)]. Do tego celu wykorzystuje się różnego rodzaju modele matematyczne, które w oparciu o zgromadzone dane i informacje aprioryczne pozwalają oszacować czasy retencji, w jakich anality opuszczą kolumnę chromatograficzną.

Do tego celu wykorzystuje się zarówno modele mechanistyczne, jak i statystyczne. Te pierwsze opierają się na podstawowych zasadach i mechanizmach związanych z procesem rozdzielania. Modele te mają na celu opisanie interakcji między cząsteczkami analitu, fazą stacjonarną i fazą ruchomą. Modele te zapewniają wgląd w podstawowe procesy fizyczne i chemiczne zachodzące podczas rozdzielania i mogą pomóc w zrozumieniu zachowania retencji związków w RP-HPLC. Modele mechanistyczne są zazwyczaj oparte na zasadach teoretycznych wyrażonych przez różnego rodzaju równania (np. model liniowej siły elucyjnej) i obserwacjach eksperymentalnych. Z kolei modele statystyczne w RP-HPLC najczęściej przyjmują formę ilościowych zależności struktura-retencja (*ang. Quantitative Structure-Retention Relationship, QSRR*), które mają na celu skorelowanie współczynników retencji związków z ich deskryptorami molekularnymi. Modele te mają na celu ustalenie matematycznego związku między cechami strukturalnymi analitów a ich retencją. Wykorzystują one metody statystyczne, w dużej mierze regresyjne, których parametry oszacowuje się przy użyciu algorytmów uczenia maszynowego [Bouwmeester et al. (2019)]. Mają one na celu zidentyfikować deskryptory molekularne, które są najsilniej skorelowane z czasami retencji. Modele te można wykorzystać do przewidywania zachowania retencyjnego nowych związków na podstawie ich informacji strukturalnych, bez potrzeby przeprowadzania szeroko zakrojonych pomiarów eksperymentalnych.

Modele mechanistyczne umożliwiają dokładniejsze zrozumienie procesu rozdzielania i mogą być wykorzystywane do optymalizacji warunków chromatograficznych w oparciu o istniejącą teorię chromatograficzną. Jednak często wymagają bardziej szczegółowych informacji o analicie (danych wstępnych) oraz mogą być trudniejsze do opracowania i walidacji. Z kolei modele statystyczne oferują bardziej praktyczne i wydajne podejście do przewidywania zachowania retencji, ponieważ opierają się na korelacjach między deskryptorami molekularnymi a czasami retencji. Można je stosować do szerokiego zakresu związków i są szczególnie przydatne do badań przesiewowych. Mogą one jednak nie uchwycić pełnej złożoności mechanizmów separacji. Są również często nieekstrapolowalne, tj. mają ograniczone zastosowanie poza zakresem związków używanych do

opracowania modelu. W praktyce w RP-HPLC często stosuje się kombinację modeli mechanistycznych i statystycznych, wykorzystując mocne strony każdego podejścia do lepszego zrozumienia, przewidywania i optymalizacji rozdzieleń chromatograficznych. W dalszej części pracy zostaną przedstawione najczęściej wykorzystywane modele stosowane w analizie dużych zbiorów danych chromatograficznych.

Najpopularniejszym modelem wykorzystywanym do predykcji czasu retencji jest regresja wektorów podpierających, inaczej wspierających czy nośnych (*ang. Support Vector Regression*, SVR). Jest ona rozszerzeniem maszyny wektorów podpierających (*ang. Support Vector Machine*, SVM) służącej do klasyfikacji, która traktuje każdą obserwację z próby jako punkt w wielowymiarowej przestrzeni cech i wyznacza hiperpłaszczyznę prawidłowo klasyfikującą te obserwacje poprzez maksymalizację marginesu między wektorami nośnymi, tj. punktami danych najbliższych hiperpłaszczyźnie. W przypadku regresji, zamiast znajdowania hiperpłaszczyzny, która w dużym stopniu rozdziela obserwacje treningowe, wprowadza się funkcję straty niewrażliwą na odchylenia rzędu $\epsilon > 0$ w celu obliczenia hiperpłaszczyzny w taki sposób, że przewidywane wartości są oddalone co najwyżej o odchylenie ϵ od wartości obserwowanych (rzeczywistych) [Zhang i O'Donnell (2020)]. Zaletami tego podejścia są nieczułość na przeuczanie modelu oraz fakt, że dobrze się uogólnia nawet przy niewielkiej liczbie obserwacji treningowych [Huang et al. (2022)]. Natomiast jego ograniczenia związane są z szybkim wzrostem wymagań obliczeniowych i pamięciowych w odniesieniu do liczby obserwacji treningowych [Haykin (2009)].

Jednymi z popularniejszych metod estymacji używanych w badaniu retencji są różne warianty regresji cząstkowych najmniejszych kwadratów (*ang. Partial Least Squares*, PLS). Regresja PLS stanowi uogólnienie modelu liniowej regresji wielorakiej. Polega na znalezieniu metodą regresji liniowej zależności między czasami retencji stanowiącymi zmienną zależną Y a nowym zestawem zmiennych, które są liniową kombinacją zmiennych niezależnych X będących zwykle deskryptorami molekularnymi (obejmującymi właściwości fizykochemiczne) badanych związków. Te nowe zmienne są obliczane w ten sposób, aby maksymalizować kowariancję między X i Y [Haddad et al. (2016); Wold et al. (2001); Abdi (2010)]. Upraszczając, regresja PLS wykrywa zmienne najbardziej związane z Y i przeprowadza na nich regresję liniową. Zaletą tego podejścia jest jej użyteczność w przypadku danych z liczbą predyktorów większą niż liczba zgromadzonych obserwacji. Ma zastosowanie również w przypadku zmiennych niezależnych, które są ze sobą skorelowane, co w przypadku deskryptorów molekularnych często ma miejsce. Natomiast głównymi niedogodnościami w stosowaniu tej metody są ryzyko przeoczenia rzeczywistych korelacji oraz zależność od skalowania predyktorów [Cramer (1993)].

Kolejną popularną metodą wykorzystywaną do badań retencji jest bayesowska regresja grzbietowa (*ang. Bayesian Ridge Regression, BRR*). Metoda ta łączy wnioskowanie bayesowskie i należącą do klasy modeli regresji liniowej regresję grzbietową, która w odróżnieniu od zwykłej regresji wielorakiej i stosowanej w niej metody najmniejszych kwadratów radzi sobie z problem skorelowania predyktorów, a co za tym idzie, osobliwością macierzy X^TX ze wzoru estymatorów współczynników $\hat{\beta} = (X^TX)^{-1}X^Ty$, gdzie X jest macierzą realizacji zmiennych objaśniających. Regresja grzbietowa radzi sobie z tą trudnością poprzez dodanie do przekątnej tej macierzy, przed jej odwróceniem, stałej dodatniej wartości λ określanej jako parametr kary, w wyniku otrzymując współczynniki regresji $\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$. Nałożenie kary λ na estymatory współczynników powoduje zmniejszenie ich wartości [Trzęsiok (2014)]. Wraz ze wzro-

stem wielkości kary maleją one do zera, ale nigdy go nie osiągają. Stąd regresja ta nie przeprowadza selekcji zmiennych. Pozwala jednak zapobiegać nadmiernemu dopasowaniu (*ang. overfitting*) i poprawia dokładność przewidywań poprzez zmniejszenie wariancji oszacowań, co wiąże się ze wzrostem błędu systematycznego, ale jest on zazwyczaj niewielki w porównaniu z redukcją wariancji. Podejście bayesowskie pozwala na uwzględnienie skwantyfikowanej wiedzy odnośnie nieznanych parametrów, a co za tym idzie oszacowanie niepewności otrzymanych wyników [Montesinos López et al. (2022)]. Głównymi niedogodnościami regresji grzbietowej są trudności w interpretowalności współczynników regresji, wynikającą z faktu, że ich oszacowania są zależne od wartości λ, oraz ustalenie rozkładów *a priori* dla parametrów, które często jest nieco arbitralne, w szczególności dla parametru kary λ.

Zbliżoną do klasycznej regresji grzbietowej metodą wykorzystywaną do predykcji retencji jest regresja Lasso (*ang. Least Absolute Shrinkage and Selection Operator regression, LASSO*). W odróżnieniu od tej pierwszej może zerować niektóre współczynniki modelu. Wynika to z jej konstrukcji, a mianowicie:

$$\hat{\beta}_{LASSO} = \arg\min_{\alpha} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_{ik} \right)^2,$$

przy ograniczeniu:

$$\sum_{k=1}^{p} |\beta_k| \leqslant s,$$

gdzie s oznacza stałą, powiązaną w sposób wzajemnie jednoznaczny z parametrem λ. Dla porównania, współczynniki z modelu regresji grzbietowej możemy zapisać w postaci identycznego warunku, z wyjątkiem ograniczenia:

$$\sum_{k=1}^{p} \beta_k^2 \leqslant s.$$

Stąd regresja Lasso pozwala na selekcję zmiennych [Tibshirani (1996)]. Jednak wybór parametru λ może mieć bardziej drastyczne skutki niż w regresji grzbietowej, w postaci wyzerowania zbyt wielu współczynników. Problematyczna może okazać się również interpretacja z uwagi na to, że współczynniki wyznaczane są w sposób dający najlepszą łączną prognozę, a nie na podstawie dokładności oszacowania i interpretacji udziału poszczególnych zmiennych predykcyjnych [Ranstam i Cook (2018)], co ma miejsce również w przypadku regresji grzbietowej. Niemniej Lasso wydaje się być użyteczną metodą w identyfikacji małego podzbioru predyktorów o najsilniejszych efektach ze zbioru wielu zmiennych predykcyjnych.

W analizie danych retencyjnych coraz częściej stosowane są sztuczne sieci neuronowe (SSN, *ang. Artificial Neural Networks, ANN*). Ich działanie w uproszczeniu ma odpowiadać działaniu biologicznych struktur nerwowych złożonych z neuronów. Każda sieć neuronowa jest opisana przez trzy główne elementy, a mianowicie charakter węzła, topologię sieci i reguły uczenia się. Charakter węzła opisuje sposób przetwarzania sygnałów przez węzeł, a więc wskazuje liczbę wejść i wyjść z danego węzła wraz z powiązanymi do nich wagami oraz funkcję aktywacji. Topologia sieci określa mechanizm, w jaki węzły są zorganizowane i połączone. Z kolei reguły uczenia opisują metody inicjowania i dostosowywania wag [Zou et al. (2009)]. Sieci neuronowe pozwalają na tworzenie odwzorowań nieliniowych i wykorzystywanie wielowymiarowych danych. Wykrywają również wszystkie możliwe interakcje pomiędzy predyktorami. Można także rozwijać

je przy użyciu wielu różnych algorytmów uczących [Tu (1996)]. Potrafią poradzić sobie z zaszumionymi czy niepełnymi danymi. Jednak korzystanie z nich wiąże się też z pewnymi problemami. Pierwszym z nich jest wymaganie dużej ilości danych próbnych. Kolejny stanowi procedura optymalizacji topologii warstw ukrytych sieci, która jest czasochłonna i komplikuje proces obliczeniowy. Następną wadą sieci neuronowych jest nieadekwatność ekstrapolacji, tzn. nie są one w stanie przewidywać wyników dla danych wejściowych leżących poza przestrzenią danych treningowych [Hossain et al. (2017)]. Kolejna niedogodność w stosowaniu SSN związana jest z ich interpretację i zrozumieniem. Trudno określić, w jaki sposób sieć neuronowa doszła do uzyskanego wyniku [Zhang et al. (2018)].

Kolejną metodą stosowaną do przewidywań retencji są lasy losowe (*ang. Random Forests, RF*). Wykorzystuje się je zarówno w problemach klasyfikacyjnych jak i regresyjnych. Stanowią one zbiór drzew decyzyjnych. Każde z nich jest zbudowane na innym podzbiorze zbioru danych treningowych, z którego obserwacje losowane są ze zwracaniem (próba bootstrapowa). Dzięki temu uzyskuje się niezależne drzewa. Każdy podział w drzewie jest wybierany jako najlepszy możliwy podział dla losowo wybranego małego podzbioru zmiennych predykcyjnych. Losowanie zmiennych objaśniających ma na celu zmniejszenie korelacji między drzewami. Uczenie kończy się, gdy liczba drzew osiągnęła zadane maksimum lub błąd w próbie testowej przestał maleć. Błąd ten to z *ang. out-of-bag error* i stanowi średni błąd predykcji na każdej próbie treningowej, przy użyciu tylko tych drzew, które nie miały jej w swojej próbie początkowej. Predykcje stanowią średnią z predykcji poszczególnych drzew [Breiman (2001)]. Zaletą lasów losowych jest ich odporność na wartości odstające i zaszumione dane. Charakteryzują się one również wysoką precyzją przewidywań i automatyczną selekcją najważniejszych cech [Genuer et al. (2010)]. Stanowią one jednak metody o dużej złożoności obliczeniowej i trudności w interpretacji [Breiman (2001)].

Innym podejściem zbliżonym do poprzedniej metody również wykorzystywanym w estymacji retencji jest tzw. wzmocnienie adaptacyjne (AdaBoost, *ang. Adaptive Boosting, AB*). Algorytm ten również wykorzystuje drzewa decyzyjne. Buduje on każde kolejne drzewo w sposób kompensujący błędy swojego poprzednika. Jest to możliwe dzięki przypisaniu wag do obserwacji, które są aktualizowane po każdej iteracji, a ich wielkości zależą od ich błędu. Stąd drzewa słabo przewidujące, nazywane słabymi uczniami, mogą skupić się na najtrudniejszych przykładach. Każde drzewo dokonuje predykcji, a ostateczna prognoza opiera się na ważonej kombinacji tych indywidualnych predykcji. Wagi w tej kombinacji przypisane do każdego słabego ucznia są oparte na jego rezultatach na danych treningowych [Wang (2012)]. Zalety i wady tego podejścia są zbieżne z tymi dla lasów losowych, jednak są drobne różnice. Wzmocnienie adaptacyjne jest mniej podatne na przeuczenie. Można osiągać precyzyjniejsze predykcje przy mniejszej ilości drzew. Przypisanie wagi każdemu słabemu uczniowi w algorytmie AdaBoost zapewnia wgląd we względne znaczenie różnych cech w zbiorze danych. Z drugiej strony wzmocnienie adaptacyjne jest podatniejsze na stronniczość (*ang. bias*). Ma to miejsce w przypadku, gdy słabi uczniowie są zbyt prości, ponieważ wtedy algorytm może nie uchwycić złożonych wzorców w danych [Shanmugasundar et al. (2021); Breiman (2001)].

Podobną metodą do poprzedniej jest tzw. wzmocnienie gradientowe (*ang. Gradient Boosting, GB*). Algorytm ten iteracyjnie dopasowuje prosty model regresji (proste drzewo decyzyjne, czy inaczej słabego ucznia) do reszt poprzedniego modelu. Reszty te to różnice między rzeczywistymi a przewidywanymi wartościami z poprzedniego

modelu. Idea tej metody polega na wykorzystaniu informacji zawartych w resztach do poprawy ogólnej wydajności modelu. Słabe i mocne strony tego podejścia są podobne jak w przypadku metod lasów losowych czy wzmocnienia adaptacyjnego. Wzmocnienie gradientowe wydają się być najdokładniejszym rozwiązaniem z wymienionych, dającym najprecyzyjniejsze predykcje. Jednak często wymagają bardziej starannego doboru hiperparametrów [Hastie et al. (2009); Friedman (2002)].

W tej pracy doktorskiej proponowane jest nowe podejście do opisu retencji, a mianowicie modele hierarchiczne wykorzystujące wnioskowanie bayesowskie i to im zostaną poświęcone następne podrozdziały.

1.5.1 Modele hierarchiczne

Modele mieszane stanowią rozszerzenie regresji, a ich cechą charakterystyczną jest zawieranie tzw. efektów stałych i losowych. Efekty stałe odpowiadają parametrom populacji, które są niezmienne dla każdych danych zebranych z populacji. Natomiast efekty losowe związane są z parametrami, których wartości różnią się dla poszczególnych przedstawicieli populacji (tj. dla każdego analitu). Nie ma jednak uniwersalnej reguły przydzielania rodzaju efektu do danego parametru, gdyż w jednym modelu ta sama zmienna może być uwzględniona przez efekt stały, a w innym przez losowy. Przypisanie typu efektu w dużej mierze zależy od interpretacji i kontekstu. Aczkolwiek jest ono o tyle istotne, że może prowadzić do różnych wyników w zależności od tego, jakim typem efektu opiszemy daną zmienną [Biecek (2013)].

Modele hierarchiczne (*ang. hierarchical models*) lub inaczej wielopoziomowe (*ang. multilevel models*) stanowią szczególny przypadek modeli mieszanych (*ang. mixed models*) [Koronacki i Ćwik (2008)]. Dokładniej, opisują sytuacje, w których efekty losowe są zagnieżdżone (*ang. nested*) [Pinheiro i Bates (2006); Biecek (2013)]. Słowo ,,zagnieżdżone'' odnosi się do hierarchicznej struktury danych i oznacza, że jednostki analizy niższego poziomu, jak na przykład anality, są pogrupowane w jednostki analizy wyższego poziomu, jak na przykład stopień zdysocjowania czy obecność podstawników. W modelach tego typu zmienność wynikowej zmiennej jest dzielona na różne poziomy.

Ignorowanie struktury hierarchicznej danych poprzez zastosowanie tradycyjnej regresji wielorakiej może prowadzić do niepoprawnych wniosków, jak na przykład wskazać fałszywie statystycznie istotne różnice w retencji między analitami, które w przypadku uwzględnienia grupowania stają się nieistotne. Wynika to z faktu niedoszacowania błędów standardowych, gdy zależność między poszczególnymi obserwacjami nie jest brana pod uwagę, jak ma to miejsce w konwencjonalnych testach statystycznych [Hox et al. (2017)]. Drugim skrajnym podejściem ignorującym hierarchiczną strukturę danych jest traktowanie każdej grupy związków oddzielnie, dopasowując do każdej z nich inny model regresji. Taka metodyka może być właściwa w sytuacji, gdy analizowanych grup analitów jest mało i każda z nich zawiera umiarkowanie dużą liczbę substancji albo gdy interesuje nas wyciąganie wniosków tylko na temat tych konkretnych grup analitów. Jeśli jednak traktujemy grupy związków jako próbkę losową z populacji wszystkich grup i chcemy wyciągać wnioski na temat ogólnych różnic między grupami, wtedy niezbędne jest zastosowanie pełnego podejścia wielopoziomowego [Goldstein (2011)].

Jednak modele hierarchiczne wraz ze wzrostem liczby poziomów mają tendencję do szybkiego komplikowania się. Trudność polega na oszacowaniu znacznej liczby parametrów, co często prowadzi do przekroczenia mocy obliczeniowej komputera. Zatem należy być ostrożnym przy dodawaniu kolejnych komplikacji. Szczególnie w sytuacji uwzględnia-

nia interakcji pomiędzy kowariantami z różnych poziomów. Należy je wprowadzać tylko wtedy, gdy istnieją wyraźne przesłanki merytoryczne, że wartość predyktora wyższego poziomu wyjaśnia charakter zależności pomiędzy zmienną odpowiedzi a predyktorem niższego poziomu albo gdy oszacowanie współczynnika kierunkowego charakteryzuje się dużą wariancją [Hox et al. (2017)].

W pracach będących podstawą tej rozprawy do opisu logarytmu współczynnika retencji wykorzystano nieliniowe modele hierarchiczne. Wielopoziomowa wersja nieliniowego modelu efektów mieszanych, zaproponowana przez Lindstroma i Batesa (1990) dla dwóch poziomów zagnieżdżenia i zapisana jako model dwuetapowy, została przedstawiona poniżej. Pierwszy etap tego modelu wyraża odpowiedź y_{ijk} dla k-tej obserwacji na i-tej grupie pierwszego poziomu i j-tej grupie drugiego poziomu jako:

$$y_{ijk} = f(\phi_{ijk}, x_{ijk}) + \epsilon_{ijk},$$
$$i = 1, \ldots, M,\ j = 1, \ldots, M_i,\ k = 1, \ldots, n_{ij},$$

gdzie M oznacza liczbę grup pierwszego poziomu, M_i liczbę grup drugiego poziomu w obrębie i-tej grupy pierwszego poziomu, a n_{ij} liczbę obserwacji na j-tej grupie drugiego poziomu wewnątrz i-tej grupy pierwszego poziomu. Natomiast ϕ_{ijk} stanowi wektor parametrów, x_{ijk} wektor kowariant, ϵ_{ijk} błąd wewnątrzgrupowy o rozkładzie normalnym $N(0, \sigma^2)$, a f funkcję nieliniową przynajmniej w jednym składniku ϕ_{ijk}. Drugi etap modelu wyraża ϕ_{ijk} jako:

$$\phi_{ijk} = \mathbf{A}_{ijk}\beta + \mathbf{B}_{i,jk}\mathbf{b}_i + \mathbf{B}_{ijk}\mathbf{b}_{ij},$$
$$\mathbf{b}_i \sim N(0, \Phi_1),\ \mathbf{b}_{ij} \sim N(0, \Phi_2),$$

gdzie β jest p-wymiarowym wektorem parametrów populacyjnych (związanych z efektami stałymi). Następne dwa oznaczenia związane są efektami losowymi, mianowicie $\mathbf{b}_i$ odpowiada q_1-wymiarowym niezależnym wektorom parametrów indywidualnych pierwszego poziomu z macierzą kowariancji Φ_1, a $\mathbf{b}_{ij}$ odpowiada q_2-wymiarowym niezależnym wektorom parametrów indywidualnych drugiego poziomu z macierzą kowariancji Φ_2. Z kolei $\mathbf{A}_{ijk}$, $\mathbf{B}_{i,jk}$ i $\mathbf{B}_{ijk}$ są macierzami eksperymentu, pierwsza efektów stałych, a dwie pozostałe efektów losowych zależnych od grup odpowiednio pierwszego i drugiego poziomu. Zakłada się, że błędy wewnątrzgrupowe ϵ_{ijk} są niezależne od efektów losowych [Pinheiro i Bates (2006)].

Metodą zwykle wykorzystywaną do oszacowania parametrów tego typu modeli jest estymacja największej wiarygodności. Polega ona na skonstruowaniu funkcji wiarygodności odpowiadającej zaobserwowanym danym, zależnym od szukanych parametrów. Następnie funkcja ta jest marginalizowana z uwagi na występowanie efektów losowych, gdyż zazwyczaj mają one rozkład wielowymiarowy, a ich wartości nie są bezpośrednio obserwowane. Dalej wyznacza się takie wartości parametrów modelu, dla których zmarginalizowana funkcja wiarygodności osiąga największą wartość. Metoda ta pozwala również wyznaczyć asymptotyczne błędy standardowe, które można wykorzystać do przetestowania istotności badanych parametrów lub wyznaczenia dla nich przedziałów ufności. Alternatywnymi metodami estymacji parametrów modeli wielopoziomowych są: *profile likelihood*, odporne błędy standardowe (*ang. robust standard errors*), bootstrap czy metody bayesowskie [Hox (2002)]. Pierwsza z nich wybiera wartość parametru z pewnego zbioru stałych. Przy każdej rozważanej wartości, przeprowadzana jest maksymalizacja funkcji wiarygodności poprzez optymalizację wszystkich innych parametrów. Następnie bada się zależność między wartością parametru a uzyskaną dla niej maksymalną wartością funkcji wiarygodności. Większe maksimum wskazuje na bardziej

prawdopodobną wartość parametru. Niepewność predykcji można obliczyć analogicznym podejściem [Cole (2014); Mitra i Hlavacek (2019)]. W przypadku metody odpornych błędów standardowych parametr zwykle estymowany jest metodą najmniejszych kwadratów, ale jego błędy standardowe są szacowane z ominięciem założenia o pochodzeniu wartości błędów z tego samego rozkładu [Croux et al. (2004)]. Odporne błędy standardowe w metodzie Hubera-White'a są równe pierwiastkom kwadratowym elementów na przekątnej macierzy kowariancji [Freedman (2006)]. Z kolei bootstrap szacuje rozkład błędów estymacji przy pomocy wielokrotnego losowania ze zwracaniem danych z próby [Efron i Tibshirani (1994)]. Opisane metody w podstawowych wersjach wykorzystują podejście klasyczne. Jednak w rozprawie doktorskiej skupiono się na metodach bayesowskich estymacji parametrów modeli hierarchicznych. W dalszej części dogłębniej omówiono, czym wyróżnia się to podejście.

1.5.2 Podejście bayesowskie

Wnioskowanie bayesowskie wykorzystuje twierdzenie Bayesa do aktualizacji prawdopodobieństwa interesujących nas hipotez na podstawie dotychczasowej wiedzy i nowych danych eksperymentalnych. Twierdzenie to wyraża się następującym wzorem:

$$P(\theta|y) = \frac{P(\theta) \cdot P(y|\theta)}{P(y)},$$

gdzie θ odnosi się do parametrów, a y do zaobserwowanych danych. $P(\theta)$ opisuje rozkład *a priori* reprezentujący stan naszej wiedzy o parametrach przed wykonaniem eksperymentu, a $P(y|\theta)$ to funkcja wiarygodności wyrażająca prawdopodobieństwo uzyskania otrzymanych danych przy założeniu ustalonych parametrów model. Wyrażenie $P(y)$ odzwierciedla rozkład brzegowy, który nie jest zależny od parametrów i możemy go traktować jako czynnik normalizujący, zaś $P(\theta|y)$ oznacza rozkład *a posteriori*, który odzwierciedla stan naszej wiedzy po uwzględnieniu zaobserwowanych danych [Grzenda (2012)].

Podejście bayesowskie a klasyczne

W teorii wnioskowania statystycznego wyróżniamy dwa główne podejścia: klasyczne i bayesowskie. Mimo iż statystyka bayesowska ma dłuższą historię niż statystyka klasyczna, to przez lata nie była popularna. Jednak w ostatnich latach zyskuje ona uznanie coraz większej liczby naukowców [Hackenberger (2019)].

Podejścia te różnią się w kilku kluczowych kwestiach. Pierwszą z nich jest interpretacja prawdopodobieństwa. Prawdopodobieństwo w klasycznym (częstościowym) ujęciu jest miarą częstości występowania badanego zjawiska w długiej serii powtórzeń eksperymentu losowego. Z tego punktu widzenia stwierdzenie prawdopodobieństwa np. pojedynczego rozdzielenia musi być osadzone w długiej sekwencji identycznych eksperymentów. Jednak nie wszystkie doświadczenia można powtórzyć przy tych samych warunkach. Natomiast w ujęciu bayesowskim prawdopodobieństwo wyraża racjonalne przekonanie reprezentujące stan wiedzy i traktowane jest jako miara pewności. Pozwala na formułowanie twierdzeń o dostępnej wiedzy cząstkowej, zdobytej na podstawie danych, dotyczącej interesującego nas zdarzenia nieobserwowalnego lub jeszcze nieobserwowanego w sposób systematyczny [Gelman et al. (2013)]. Zatem podejście bayesowskie nie odnosi się do granicznych własności statystyk, jak to ma miejsce w przypadku

podejścia klasycznego i można je stosować zarówno w przypadkach, gdy podejście częstościowe ma zastosowanie, jak również wtedy, gdy nie ma ono sensu.

Dodatkowo w podejściu klasycznym parametry modelu interpretowane są jako nieznane stałe. Z kolei w ujęciu bayesowskim przyjmuje się, że szacowane parametry są zmiennymi losowymi. Jest to konsekwencją subiektywnej interpretacji prawdopodobieństwa i traktowaniu go jako stopnia ufności wobec różnych wartości oszacowywanych parametrów. Pozwala to na określenie rozkładu wyrażającego nasze wstępne przypuszczenie o nieznanych parametrach na zbiorze ich możliwych wartości [Grzenda (2012)].

Kolejną cechą rozróżniającą oba podejścia jest możliwość uwzględnienia w modelu informacji spoza próby. Metody klasyczne opierają się wyłącznie na informacjach płynących z próby i uniemożliwiają wzbogacenie wnioskowania o dodatkowe wiadomości, zaś metody bayesowskie pozwalają na uwzględnienie dodatkowej wiedzy *a priori* pochodzącej np. z poprzednich badań czy literatury. Dzięki temu możliwe jest uzyskanie dokładniejszych wyników, pod warunkiem że informacja *a priori* nie jest przypadkowa. W przeciwnym razie otrzymane rezultaty byłyby mało wiarygodne. Natomiast przy odpowiednio dobranych rozkładach *a priori* uzyskane wyniki będą analogiczne do tych uzyskanych metodami klasycznymi, jednak ich interpretacja będzie inna. Stąd niezwykle istotny jest właściwy ich dobór i konieczność przedstawienia ich w wynikach [Kruschke (2021)].

Dobór rozkładu *a priori*

Formalnie informacja *a priori* pozwala na uwzględnienie istotnych dla analizowanego problemu wiadomości. W praktyce często stanowi środek do stabilizacji wnioskowań w złożonych, wielowymiarowych modelach, a czasem wręcz niedogodność, służąc jedynie jako ,,katalizator wyrażania niepewności za pomocą twierdzenia Bayesa'' [Gelman et al. (2017)].

Teoretycznie ustalenie rozkładu *a priori* powinno poprzedzać model, jednak w praktyce często jest on wybierany w odniesieniu do funkcji wiarygodności, tzn. zbierane są dane i tworzony jest model zawierający różne nieznane parametry, a następnie ustalane są rozkłady *a priori*, bez których wnioskowanie bayesowskie jest niemożliwe. Ich użyteczność w ramach danej analizy zależy od tego, jak oddziałują one na założony model w kontekście rzeczywistych obserwowanych danych. Należy również pamiętać, że rozkład *a priori* dobrze dobrany w jednym scenariuszu, w drugim może być problematyczny. Zatem w celu dogłębnej analizy należy wyjść poza standardowy schemat, w którym rozkład *a priori* wybiera się bez odniesienia do danych i procesu generowania danych [Gelman et al. (2017)].

Rozkłady *a priori* wyrażają stopień przekonania co do różnych wartości szacowanych parametrów. W przypadku braku wiarygodnej wiedzy *a priori* lub po prostu chęci ograniczenia jej wpływu na wynik, wykorzystuje się tak zwane nieinformacyjne rozkłady *a priori*. Niemożliwe jest zdefiniowanie rozkładów *a priori*, które by w żaden sposób nie oddziaływały na funkcję wiarygodności, jednak te nieinformacyjne wpływają na nią najmniej jak to możliwe. Przykładem takich rozkładów są rozkłady jednostajne czy rozkłady *a priori* Jeffreysa. Z kolei w celu uwzględnienia w modelu informacji mających zastosowanie do ogólnej klasy problemów, ale bez pełnego wykorzystania wiedzy specyficznej dla danego zagadnienia, wykorzystuje się słabo informacyjne rozkłady *a priori*. Nie zawierają one pełnych informacji, stąd nie stanowią idealnego rozwiązania, ale mogą być dobrym punktem wyjścia, gdyż zapewniają pewną użyteczną regularyzację.

Estymacja parametrów modelu

Wyznaczanie rozkładów *a posteriori* w sposób analityczny bywa trudne, a w niektórych przypadkach nawet niemożliwe. Wspomniane zaniedbanie w stosowaniu podejścia bayesowskiego wynikało z ograniczoności metod numerycznych. Rozwój metod symulacyjnych pozwolił jednak na wzrost zainteresowania tym podejściem w ostatnich latach.

Obecnie w estymacji bayesowskiej stosowane są metody Monte Carlo oparte na łańcuchach Markowa (ang. Markov chain Monte Carlo, MCMC). Polegają one na generowaniu ergodycznego łańcucha Markowa, który po upływie odpowiednio długiego czasu (wielu iteracjach) osiągnie rozkład stacjonarny, w podejściu bayesowskim określanym jako rozkład *a posteriori*. W praktyce niestety trudno określić ile iteracji łańcucha Markowa wystarczy, aby tworzony przez nie zbiór stanów tworzył rozkład stacjonarny. Problem z określeniem, czy rozkład stacjonarny został osiągnięty, nazywamy problemem zbieżności łańcucha.

Kolejnym czynnikiem, jaki należy wziąć pod uwagę, jest fakt, że algorytm symulacyjny wymaga przyjęcia pewnych wartości inicjujących, których wpływ maleje ze wzrostem długości symulacji. Stąd, aby zminimalizować skutek oddziaływania tych wartości i z uwagi, że początkowe iteracje nie dostarczają wiarygodnych informacji o rozkładzie *a posteriori*, standardową praktyką jest ich odrzucenie (tzw. *burn-in* lub *warmup*) [Brooks et al. (2011)].

Następną rekomendacją jest symulacja co najmniej trzech niezależnych łańcuchów inicjowanych różnymi wartościami początkowymi. Mieszając skrócone o początkowe iteracje symulacje z kilku łańcuchów, ograniczamy problem związany z autokorelacją wewnątrz łańcuchów, która może skutkować złą estymacją wariancji.

Problem zbieżności łańcucha próbuje się ocenić za pomocą graficznej analizy wykresu ,,tropu'' (*ang. traceplot*) parametrów. Przedstawia on wartości danego parametru w kolejnych iteracjach algorytmu. Gdy łańcuchy są zbieżne, wykres przybiera formę określaną jako ,,gruba, włochata gąsienica'', w przeciwnym wypadku przypomina on bardziej ,,wijącego się węża'' [Lunn et al. (2013)]. Również wykresy autokorelacji mogą pomóc w ocenie zbieżności. W przypadku zbieżności łańcuchów powinniśmy obserwować zmniejszające się wartości autokorelacji [Ntzoufras (2011)]. Kolejną miarą pozwalającą na ocenę zbieżności jest statystyka ,,potencjalnej redukcji skali'' (*ang. potential scale reduction statistic*, $\hat{R}$). Czynnik ten mierzy stosunek średniej wariancji próbek w każdym łańcuchu do wariancji połączonych próbek z różnych łańcuchów. Jeśli wszystkie łańcuchy są zbieżne, to $\hat{R}$ będzie równe jeden. Uznaje się, że łańcuchy nie osiągnęły zbieżności, gdy $\hat{R}$ jest większe niż 1.1 [Kubacki (2014); Gelman i Rubin (1992)]

Najbardziej znanymi algorytmami metod Monte Carlo opartych na łańcuchach Markowa są algorytm Metropolisa-Hastingsa oraz jego szczególny przypadek: próbnik Gibbsa. Jednak istnieją bardziej efektywne algorytmy, jednym z nich jest wykorzystywany w programie Stan próbnik ,,bez zawracania'' (*ang. No-U-Turn Sampler, NUTS*), który jest rozszerzeniem hybrydowej czy inaczej hamiltonowskiej metody Monte Carlo (*ang. Hybrid/Hamiltonian Monte Carlo, HMC*) [Hoffman i Gelman (2014)].

Diagnostyka modelu

Sprawdzanie modelu jest kluczowym elementem każdej analizy statystycznej. Niezwykle trudno uwzględnić w rozkładzie prawdopodobieństwa całą posiadaną wiedzę

na temat badanego problemu, stąd należy zbadać, jakich aspektów rzeczywistości model nie ujmuje. W statystyce bayesowskiej wyróżniamy co najmniej trzy sposoby na diagnostykę modelu.

1. Badanie wrażliwości wnioskowań na uzasadnione zmiany w rozkładzie *a priori* i funkcji wiarygodności.

   Nawet gdy mamy pewność, że dane nie są sprzeczne z przyjętym modelem, to nie wystarcza, by wzbudzić pełne zaufanie do ogólnych wniosków merytorycznych. Wynika to z faktu, że mogą istnieć inne rozsądne modele, które zapewnią równie dobre dopasowanie, ale prowadzą do innych wniosków. Stąd analiza wrażliwości może być wykorzystana do oceny wpływu alternatywnych analiz na wnioskowania płynące z rozkładu *a posteriori*.

2. Merytoryczne sprawdzenie, czy wnioski płynące z rozkładu *a posteriori* są uzasadnione.

   W każdym problemie badawczym znajdują się informacje, które nie są uwzględnione w modelu (ani w rozkładzie *a priori*, ani w funkcji wiarygodności) czy to z uwagi na obiektywizm, czy po prostu z wygody. Jeśli te dodatkowe informacje sugerują, że wnioski z otrzymanego rozkładu *a posteriori* są fałszywe, to należałoby stworzyć dokładniejszy model albo chociaż mieć świadomość słabych jego punktów.

3. Sprawdzenie, czy model dopasowuje się do danych.

   Model możemy sprawdzić, wykorzystując tzw. walidację zewnętrzną (*ang. external validation*). Metoda ta opiera się na porównaniu prognozowanych przez model przyszłych danych i zebranych nowych danych. Jednak istnieją metody aproksymujące walidację zewnętrzną przy użyciu danych, które już posiadamy. Jedną z tych metod oceny jest badanie tzw. predykcyjnego rozkładu *a posteriori* (*ang. posterior predictive distribution*) [Gelman et al. (1996)]. Rozkład ten wyraża się wzorem:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)\, d\theta \\ &= \int p(\tilde{y}|\theta, y) \cdot p(\theta|y)\, d\theta, \end{aligned}$$

   gdzie $\tilde{y}$ oznacza przewidywane nieznane wartości obserwowalne. Zatem wspomniana metoda polega na symulowaniu wartości z predykcyjnego rozkładu *a posteriori* i porównaniu ich z obserwowanymi danymi. Jeśli model jest dobrze dopasowany, to te wartości są podobne. Jakiekolwiek systematyczne różnice między symulacjami a danymi wskazują na potencjalne wady modelu. W praktyce wiele powszechnie stosowanych metod sprawdzania modelu, takich jak testy dla wartości odstających, wykresy reszt i wykresy normalne, można interpretować jako ,,predykcyjne testy bayesowskie *a posteriori*'' (*ang. Bayesian posterior predictive checks*) [Gelman (2003)].

1.5.3 Projektowanie eksperymentów

Praktycy coraz częściej szukają skutecznych podejść, do wyznaczenia najbardziej selektywnych kolumn i metod chromatograficznych, które pozwoliłyby rozwiązać problemy separacji w możliwie najkrótszym czasie [Gritti (2021)]. Często prowadząc badania

eksperymentalne, musimy podjąć decyzję, jakie warunki chromatograficzne wybrać, zanim uzyskamy jakiekolwiek dane. Kolejną niedogodnością jest ograniczoność zasobów [Chaloner i Verdinelli (1995)]. Stąd projektowanie eksperymentów (*ang. design of experiments, DOE)* ma coraz większe znaczenie. Polega ono na zaproponowaniu matematycznego/statystycznego modelu opisującego retencję, który następnie wykorzystywany jest *in silico* do wspomagania wyszukiwania optymalnych warunków chromatograficznych do rozdzielenia najbardziej krytycznej pary analitów w mieszaninie próbek. Modele używane w tego rodzaju analizach mają charakter czysto statystyczny bądź opierają się na podstawach chromatografii. Te pierwsze nie uwzględniają elementarnych zasad chromatografii cieczowej i sprawdzają się przy rozwiązywaniu bardzo złożonych problemów, dla których nie istnieją oczywiste modele analityczne. Z kolei te drugie opierają się na empirycznych wyrażeniach dla modeli retencji (jak na przykład liniowa moc rozpuszczalnika) oraz na podstawowych zależnościach chromatografii. Stanowią one solidne narzędzia i dobrze nadają się do przewidywania czasów retencji w metodach RP-HPLC [Gritti (2021)].

Przy projektowaniu eksperymentów określa się tzw. funkcję użyteczności (*ang. utility function*), która pozwala na podjęcie najlepszej decyzji doboru eksperymentu. W książce Lindley (1972) autor proponuje, aby specyfikacja funkcji użyteczności odzwierciedlała cel eksperymentu, by traktować wybór projektu jako problem decyzyjny i wybrać taki projekt, który maksymalizuje oczekiwaną użyteczność. Tutaj należy pamiętać, że projekt, który optymalizuje estymacje, niekoniecznie będzie optymalizował predykcje.

Do problemu wyboru odpowiedniego eksperymentu można podejść w sposób sekwencyjny. W przypadku modeli nielinowych funkcja użyteczności *a posteriori* zależy od nieznanych parametrów i tu powinno obserwować się korzyści z sekwencyjnego doboru eksperymentów [Chaloner i Verdinelli (1995)].

Rozdział 2
Założenia i cele badawcze

Celem niniejszej rozprawy było opracowanie bayesowskich modeli hierarchicznych opisujących retencję analitów w wysokosprawnej chromatografii cieczowej.

Modele zostały opracowane w oparciu o dane retencyjne dużej liczby analitów zebrane z wykorzystaniem zarówno izokratycznej, jak i gradientowej chromatografii cieczowej. Zadaniem opracowanych modeli była charakteryzacja retencji analitów oraz własności fazy stacjonarnej i ruchomej. Takie modele są niezbędne do opracowania procedury decyzyjnej, która będzie praktycznie użyteczna do przewidywania czasów retencji i szerokości piku chromatogramu, a w konsekwencji w określeniu warunków prowadzących do uzyskania żądanego rozdzielenia.

Cele szczegółowe pracy obejmowały:

1. Opracowanie modelu opisującego retencję analitów w warunkach izokratycznych z wykorzystaniem masy molowej oraz grup funkcyjnych jako predyktorów.

2. Opracowanie modelu opisującego retencję analitów w warunkach izokratycznych z wykorzystaniem $\log P$ i pK_a jako predyktorów.

3. Opracowanie bayesowskiego modelu hierarchicznego opisującego retencję analitów w warunkach gradientowych obejmujących różne wartości pH, rodzaje modyfikatora organicznego, temperatury i czasy trwania gradientu z wykorzystaniem $\log P$, pK_a oraz grup funkcyjnych jako predyktorów.

Rozdział 3
Metodyka

Podstawą tej rozprawy są trzy artykuły, w których opisane zostały matematyczne modele retencji. Te opisujące elucję izokratyczną, przedstawione w publikacjach A i B, stanowią wstęp metodologiczny do bardziej złożonego problemu, a mianowicie opisu elucji gradientowej, zaprezentowanego w artykule C. W tym rozdziale zostaną przedstawione najważniejsze założenia stosowanej metodologii. Szczegółowy opis zawarty jest w poszczególnych artykułach.

3.1 Modele

Czas retencji t_R wyznaczono na podstawie równania:

$$\int_0^{t_R - t_0 - t_e} \frac{dt}{t_0 \cdot k_i(t)} = 1. \tag{3.1}$$

Jego wartość przybliżono wykorzystując metodę trapezów. Dokładniej, algorytm sumuje pola trapezów, dopóki nie osiągnie wartości 1. To pozwala na wskazanie górnej granicy całkowania, a co za tym idzie czasu retencji t_R.

We wszystkich modelach do opisu krzywych zależności między logarytmem współczynnika retencji $\log k$ a zawartością modyfikatora organicznego φ wykorzystano nieliniowe równanie zaproponowane w pracy Neue et al. (2001):

$$\log k = \log k_w - \frac{S_1 \cdot \varphi}{1 + S_2 \cdot \varphi}, \tag{3.2}$$

gdzie $\log k_w$ oznacza logarytm współczynnika retencji w czystej wodzie, S_1 stałą opisującą zmiany retencji wraz ze składem rozpuszczalnika, a S_2 współczynnik krzywizny. Parametry te są różne w zależności od analitu. W dwóch pierwszych pracach A i B równanie (3.2) zostało sparametryzowane przy użyciu logarytmu współczynnika retencji w 100% acetonitrylu:

$$\log k_a = \log k_w - \frac{S_1}{1 + S_2}.$$

Natomiast w przypadku ostatniego artykułu C, gdzie dane eksperymentalne były znacznie bogatsze, logarytm współczynnika retencji uzależniono jeszcze od stopnia dysocjacji, modyfikatora organicznego, wartości pH i temperatury. Odnoszą się do nich odpowiednio indeksy r, m, t i b. W celu wskazania, które z parametrów zmieniają się wraz z analitami wprowadzono jeszcze indeks i, a indeks j wskazuje na zależność od

zawartości modyfikatora organicznego:

$$\begin{aligned}\log k_{r,m,b,t,i,j} = &\log k_{w_{r,i}} - \frac{S_{1_{r,m,i}} \cdot (1 + S_{2_m}) \cdot \varphi_j}{1 + S_{2_m} \cdot \varphi_j} + d\log kT_i \frac{T_t - 25}{10} + \\ &+ |chargeA_{r,i}| \cdot apH_A \cdot (pH_{m,b,t,j} - 7) + \\ &+ |chargeB_{r,i}| \cdot apH_B \cdot (pH_{m,b,t,j} - 7).\end{aligned}$$

W powyższym wzorze S_1, S_2 mają znaczenie jak poprzednio, $\log k_w$ oznacza logarytm współczynnika retencji w czystej wodzie w $25°C$ i $pH = 7$, natomiast $d\log kT_i$ reprezentuje zmianę wartości parametru $\log k_w$ wraz ze wzrostem temperatury o $10°C$. Wreszcie $chargeA_{r,i}$ i $chargeB_{r,i}$ oznaczają stan ładunku odpowiednio dla anionów i kationów, a apH efekt pH na ich retencję. W przypadku tej pracy nie wprowadzono parametru $\log k_a$, a parametry $\log k_{w_{r,i}}$, $S_{1_{r,m,i}}$ obliczano na podstawie $\log k_{w,N}$, $S_{1,a,N}$, $S_{1,m,N}$ (czyli $\log k_w$, $S_{1,a}$ i $S_{1,m}$ dla formy neutralnej analitu, gdzie indeksy a i m odnoszą się do modyfikatora organicznego odpowiednio acetonitrylu i metanolu) oraz parametrów $d\log k_w$, $dS_{1,a}$, $dS_{1,m}$ odzwierciedlających różnicę odpowiednio w $\log k_w$, $S_{1,a}$, $S_{1,m}$ pomiędzy formą neutralną a zjonizowaną. Ponadto założono liniową zależność pomiędzy wartościami pK_a a zawartością modyfikatora organicznego φ.

Modele hierarchiczne dzięki swojej strukturze (występowaniu dwóch rodzajów efektów), pozwalają na uwzględnienie podobieństwa między analitami, przy jednoczesnym indywidualnym opisie każdego z nich. Podobieństwo jest konsekwencją użycia tego samego równania opisującego zależność logarytmu współczynnika retencji $\log k$ od zawartości modyfikatora organicznego φ, oraz założenia, że specyficzne dla analitu parametry tego równania pochodzą ze wspólnego rozkładu.

Pierwszy poziom w hierarchii wyraża obserwowany czas retencji t_R (jak w przypadku ostatniej pracy C) lub logarytm współczynnika retencji $\log k$ (jak w przypadku publikacji A i B) jako wartość pochodzącą z rozkładu normalnego lub t Studenta. Średnie tych rozkładów (parametry położenia) są wartościami czasu retencji t_R lub logarytmu współczynnika retencji $\log k$ wyznaczonymi z równań (3.1) i (3.2), parametry skali wyrażone są przez σ, a w przypadku rozkładu t Studenta niezbędne jest jeszcze wskazanie liczby stopni swobody v.

Natomiast drugi poziom w hierarchii opisuje zależność pomiędzy predyktorami a wektorem specyficznych dla analitu parametrów. W publikacjach A i B jest to wektor $(\log k_w, \log k_a, S_2)$, a w publikacji C wektor $(\log k_{w,N}, S_{1,m,N}, S_{1,a,N})$. Założono, że parametry te są ze sobą skorelowane i pochodzą z wielowymiarowego rozkładu normalnego lub z wielowymiarowego rozkładu t Studenta. Średnie tych rozkładów stanowią kombinacje liniowe predyktorów, a parametry skali wyrażone są przez macierze kowariancji Ω postaci:

$$\Omega = \begin{bmatrix} \omega_{\log k_w} & 0 & 0 \\ 0 & \omega_{\log k_a} & 0 \\ 0 & 0 & \omega_{\log S_2} \end{bmatrix} \cdot \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & \rho_{2,2} & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & \rho_{3,3} \end{bmatrix} \cdot \begin{bmatrix} \omega_{\log k_w} & 0 & 0 \\ 0 & \omega_{\log k_a} & 0 \\ 0 & 0 & \omega_{\log S_2} \end{bmatrix},$$

gdzie macierz środkowa to macierz korelacji, a macierze diagonalne (pierwsza i ostatnia) wyrażają wartości skali dla odpowiednich parametrów.

3.2 Rozkłady *a priori*

We wszystkich modelach zastosowano słabo informacyjne rozkłady *a priori*. Dobierano je w taki sposób, by były zgodne z literaturową wiedzą odnośnie retencji analitów

w układzie RP-HPLC. W przypadku pierwszych artykułów (A i B) rozkłady *a priori* parametrów specyficznych dla analitów zostały wyznaczone na podstawie przybliżonego oszacowania indywidualnych parametrów metodą najmniejszych kwadratów przy pomocy równania (3.2) przy założeniu, że $S_2 = 2$. Metoda regresji liniowej pozwoliła na wyznaczenie linii regresji wyrażających zależność parametrów specyficznych dla analitów od predyktorów. Za średnie w rozkładach *a priori* przyjęto parametry tych prostych, a za skale ich odchylenia standardowe.

Założono, że wartości typowe parametrów czy współczynniki przy predyktorach pochodzą z rozkładów normalnych. Natomiast parametry modelu związane ze skalą przyjęto jako wartości z rozkładów półnormalnych, żeby uniknąć wartości ujemnych.

Wszystkie modele zostały zaimplementowane w programie R sprzężonym z programem Stan, a ich skrypty zamieszczono w serwisie GitHub pod adresami:

- `https://github.com/akamedulska/izo_1026_groups`,
- `https://github.com/akamedulska/izo_1026_logP`,
- `https://github.com/akamedulska/lc-ms`.

Rozdział 4
Wyniki

4.1 Praca badawcza A

Artykuł pt. ,,Application of Bayesian Multilevel Modeling in the Quantitative Structure-Retention Relationship Studies of Heterogeneous Compounds'' prezentuje model hierarchiczny oceniający wpływ masy molowej i grup funkcyjnych na retencję. Wykorzystano tu zarówno ilościowe zależności struktura-retencja, teorię chromatograficzną w postaci równania Neue, jak i podejście bayesowskie. To ostatnie pozwala na kwantyfikację niepewności, co w przypadku modeli opartych na zależnościach struktura-retencja z ograniczoną ilością danych ma kluczowe znaczenie.

Model zbudowano na podstawie danych chromatograficznych 1026 analitów mierzonych w warunkach izokratycznych. Dane te są ogólnodostępne na stronie internetowej: `https://www.retentionprediction.org/hplc/database/index.php`. Informacje o grupach funkcyjnych każdego związku uzyskano dzięki programowi Checkmol, a masę molową dzięki programowi ACD/Labs.

Zaproponowany model opisuje retencję analitów w oparciu o równanie Neue oraz ilościowe zależności struktura-retencja dla każdego z parametrów tego równania. Dodatkowo opisano wpływ poszczególnych grup funkcyjnych, zakładając ich podobny efekt. Wpływ podstawników czy masy molowej można łatwo odnieść do hipotetycznej sytuacji badania, jak zmieniłaby się retencja każdego związku, gdyby jego masa molowa uległa zmianie lub jedna grupa funkcyjna została zastąpiona drugą. Wartość parametru $\log k_w$ okazuje się większa dla analitów z wyższą masą molową (w przybliżeniu o 2,6 na każde 100 g/mol różnicy). Z kolei wpływ masy molowej na parametr $\log k_a$ jest prawie sześć razy mniejszy niż na parametr $\log k_w$, a dla parametru $\log S_2$ jest on znikomy. W przypadku grup funkcyjnych, obecność każdej z nich obniża parametr $\log k_w$ średnio o 0,5. Jednak współczynnik zmienności dla podstawników jest bardzo duży, co świadczy o silnej zmienności wpływu grup funkcyjnych na retencję w fazach ruchomych bogatych w wodę. Jest to prawdopodobnie związane z szeroki zakresem interakcji zachodzących między podstawnikami w tym środowisku, jak np. wiązania wodorowe czy oddziaływania dipol-dipol. Wpływ grup funkcyjnych na parametr $\log k_a$ jest z kolei mniejszy i ma mniejszą zmienność. Zastosowana metodyka, określająca wartości parametrów dla typowej grupy funkcyjnej, tym samym pozwala na oszacowanie wpływu na retencję dla grup funkcyjnych, które nie wystąpiły w danych.

Model pozwolił również wskazać wpływ poszczególnych grup funkcyjnych na retencję. Wykazano, że największy efekt na parametr $\log k_w$ wywołują następujące grupy: czwartorzędowa sól amoniowa, trzeciorzędowa amina alifatyczna i sulfonamidowe grupy funkcyjne (terminologia z programu Checkmol). Wynika to prawdopodobnie z ich

zjonizowania w pH fazy ruchomej użytej w eksperymentach (pH = 2,66). Zgodnie z oczekiwaniami, parametr $\log k_w$ jest mniejszy dla struktur alkenowych i aromatycznych. Czwartorzędowe sole amoniowe, aminy alifatyczne (pierwszorzędowe, drugorzędowe i trzeciorzędowe), grupy aromatyczne z grupą iminową oraz pochodne guanidyny należą do podstawników o najwyższym wpływie na parametr $\log k_a$ i prawdopodobnie są również zjonizowane w fazach ruchomych bogatych w acetonitryl. Najmniejszy efekt zaobserwowano dla kwasów karboksylowych. Z kolei wpływ grup funkcyjnych na $\log S_2$ jest niewielki. Kilka grup funkcyjnych (np. grupa sulfonamidowa) wykazuje odmienną retencję w fazach ruchomych bogatych w wodę niż w fazach ruchomych bogatych w acetonitryl. Wpływy grup funkcyjnych, które nie są powszechne w analizowanym zbiorze danych (grupy iminowe, hemiacetalowe oraz hydrazony) są przewidywane z dużą niepewnością. Te szerokie przedziały niepewności wskazują na brak wiedzy o tych parametrach poza tą wynikającą z ich podobieństwa do innych grup funkcyjnych.

Model pozwala przewidywać współczynniki retencji/czas retencji dla nowego analitu spoza zbioru danych, tego samego analitu w nowych warunkach chromatograficznych przy dostępie do wszystkich danych eksperymentalnych lub analitu z inną grupą funkcyjną. Do wizualizacji niepewności predykcji zaproponowano nowy typ wykresu nazwany chromatogramem niepewności, przedstawiający estymowane funkcje gęstości. Można go interpretować następująco: jeśli pewne piki na chromatogramie niepewności nie nakładają się, jest mało prawdopodobne, aby te związki miały podobny współczynnik retencji/czas retencji.

Analizując model, wykazano, że informacje płynące z grup funkcyjnych i parametrów na poziomie populacji mają ograniczoną użyteczność praktyczną ze względu na dużą niepewność uzyskanych na ich podstawie predykcji. Niepewność ta jest jednak skończona, co sugeruje, że uzyskano pewne informacje i że pewne wnioski co do tego, czy separacja analitów jest możliwa, można wyciągnąć nawet bez żadnych wstępnych danych. Co więcej, nawet pojedynczy eksperyment zwykle daje dużo informacji, i tym samym pozwala zmniejszyć niepewność predykcji.

4.2 Praca badawcza B

Artykuł pt. ,,Statistical analysis of isocratic chromatographic data using Bayesian modeling'' opisuje bayesowski model hierarchiczny stworzony na podstawie danych chromatograficznych uzyskanych w warunkach izokratycznych, który w celu predykcji wykorzystuje informacje o $\log P$ i pK_a związków.

Do konstrukcji przedstawionego w pracy modelu wykorzystano dane izokratyczne 1026 analitów dostępne na stronie internetowej: `https://www.retentionprediction.org/hplc/database/index.php`, natomiast dane o predyktorach uzyskano z pomocą programu ACD/Labs.

W publikacji opisany jest przebieg analizy tych samych danych, co w pracy A. Punktem wyjścia był model, który uwzględniał informacje tylko o jednym predyktorze, jakim była lipofilowość wyrażona przez $\log P$. Wstępna wersja modelu wykazała, że anality grupują się w dwa klastry. W celu wyjaśnienia tego fenomenu w kolejnych wersjach modelu założono, że specyficzne dla analitu parametry $(\log k_w, \log k_a, S_2)$ nie pochodzą już z wielowymiarowego rozkładu t Studenta, a z rozkładu mieszanego, dokładniej z dwóch wielowymiarowych rozkładów t Studenta. Model ten pozwolił wyliczyć, prawdopodobieństwo przynależności do określonego klastra. Dla części analitów to prawdopodobieństwo wynosiło około 50%. Ta obserwacja wydawała się być spowodowana różnym stopniem zdysocjowania związków. W związku z tym w kolejnym modelu dodano informacje o pK_a analitów. Dokonano tego poprzez wyznaczenie stopnia zdysocjowania każdego związku w pH = 2.66, odpowiadającym pH fazy ruchomej użytej w eksperymentach.

Uzyskane parametry modelu okazały się potwierdzać doniesienia literaturowe, jak np. te, że różnica w logarytmach współczynnika retencji w czystej wodzie $\log k_w$ lub w 100% acetonitrylu $\log k_a$ między formę neutralną a zjonizowaną analitu wynosi około 1, pomimo że ta informacja nie została uwzględniona w modelu. Różnica między $\log k_w$ i $\log k_a$ odzwierciedla całkowitą energię swobodną transferu analitu z wody do acetonitrylu i okazała się ona większa dla analitów z większą wartością $\log P$, co wskazuje na większą preferencję substancji lipofilowych do acetonitrylu. Model wykazał również małą różnicę we współczynniku krzywizny w zależności od stopnia zdysocjowania. Wskazuje to, że krzywe zależności logarytmu współczynnika retencji $\log k$ od zawartości modyfikatora organicznego φ dla frakcji zjonizowanych są bardziej spłaszczone.

Mimo że dane i model skonstruowany na ich podstawie nie wyjaśniają w pełni mechanizmu absorpcji, to pozwalają na predykcję retencji w sytuacji dostępu do różnych informacji (takich jak predyktory czy dane eksperymentalne). Nieuwzględnienie wiedzy o pK_a analitów prowadzi do mniej precyzyjnych przewidywań, wynikających z faktu, że przynależność do klastra musi zostać przewidziana z danych.

Zaproponowany w pracy model opisuje najważniejsze cechy danych chromatograficznych i może być użyty do przewidywania współczynnika retencji k dla nowych analitów, które są podobne do tych wykorzystanych do budowy modelu oraz analizowanych w tych samych warunkach chromatograficznych. Dla nowych analitów przewidywania są mało precyzyjne. Dokładność przewidywań poprawia się wraz z dodaniem informacji z kolejnych eksperymentów. Zaproponowany model propaguje niepewność i dzięki temu może być wykorzystany do podejmowania decyzji w warunkach niepewności. W pracy wskazano, że przewidywania oparte tylko i wyłącznie o informacje o $\log P$ i pK_a związku nie są wystarczająco dokładne, by mogły być wykorzystane w praktyce. Jednak uwzględnienie danych z chociaż jednego eksperymentu zmniejsza niepewność predykcji

na tyle, aby jej wynik był użyteczny.

4.3 Praca badawcza C

Artykuł pt. ,,Toward the General Mechanistic Model of Liquid Chromatographic Retention'' prezentuje bayesowski model hierarchiczny zbudowany na podstawie danych gradientowych zebranych dla różnych wartości pH, typów modyfikatora organicznego, temperatury oraz czasu trwania gradientu.

Dane chromatograficzne pochodziły z 187 analitów mierzonych w warunkach gradientowych podczas 84 różnych eksperymentów wykonanych w Katedrze Biofarmacji i Farmakodynamiki Wydziału Farmaceutycznego Gdańskiego Uniwersytetu Medycznego. Informacje o grupach funkcyjnych związków uzyskano na podstawie programu Checkmol, a o $\log P$ i pK_a na podstawie programu ACD/Labs.

Model opisany w pracy został zbudowany w oparciu o teorię chromatografii gradientowej, a rozkłady *a priori* dla parametrów określono na podstawie wiedzy literaturowej. Oszacowane w pracy parametry modelu na ogół wykazują zgodność z dostępną wiedzą. Najbardziej zaskakujące różnice między rozkładami *a priori* (uzyskanymi z literatury) a rozkładami *a posteriori* zaobserwowano dla parametrów dS_1a, S_2 i τ. Parametry dS_1a opisują retencję anionów (dS_1aA) i kationów (dS_1aB) w fazie ruchomej bogatej w acetonitryl. Parametry te mają różne znaki, co sugeruje różnicę w retencji pomiędzy kwasami i zasadami. Ta różnica jest mniej widoczna dla metanolu. Parametry S_2 są wyższe dla metanolu i niższe dla acetonitrylu niż oczekiwano. Wiadomo, że faza stacjonarna zmienia swoje właściwości w zależności od pH fazy ruchomej lub typu buforu. W tej pracy zmiana ta została określona ilościowo jako wartość nachylenia zależności między $\log k_w$ a pH. Nachylenie to okazało się ujemne dla anionów i dodatnie dla kationów. Efekt ten jest prawdopodobnie spowodowany kombinacją różnych mechanizmów związanych z pH fazy ruchomej, takich jak obecność powierzchniowych grup silanolowych i tworzenie się par jonowych ze składnikami buforu. Z kolei wartości parametrów τ okazały się dużo większe, niż przypuszczano. Mówią one o zmienności między analitami dla parametru α, stanowiącego współczynnik kierunkowy w równaniu liniowym opisującym zależność pK_a od zawartości modyfikatora organicznego. W pewnym stopniu może to być konsekwencją niepoprawnej identyfikacji niektórych analitów, wynikającej np. z błędnych wartości predyktorów (tj. pK_a, $\log P$, ładunków i grup).

Zaobserwowano również silne korelacje między poszczególnymi parametrami odpowiadającymi obojętnym formom analitów: między $\log k_w$ a S_1 w metanolu, między $\log k_w$ a S_1 w acetonitrylu, między S_1 w metanolu a S_1 w acetonitrylu. Dodatkowo wartości parametru α dla metanolu i acetonitrylu są silnie skorelowane, co implikuje wzajemną informację między zmiennymi. Dzięki istnieniu silnych korelacji można zdobyć wiedzę o jednym parametrze, znając wartość innego parametru (np. wiedza o $\log k_w$ zawęża zakres możliwych wartości S_1 dla formy neutralnej konkretnego analitu). To spostrzeżenie znajduje potwierdzenie w praktyce, ponieważ jeden przebieg badania gradientowego (zakładając zawartość metanolu lub acetonitrylu jako pojedynczą zmienną projektową) zwykle dostarcza wielu informacji na temat retencji.

Uzyskane z modelu parametry określone na poziomie populacji podsumowują zachowanie typowego analitu i zawierają informacje wymagane do przewidywania retencji nowych analitów.

Rozdział 5
Dyskusja wyników

W pracy zaproponowano modele, które pozwalają na bezpośrednią i łatwą interpretację wszystkich ich parametrów oraz na ujednolicony opis całego zbioru danych. To podejście wyróżnia się na tle klasycznych metod analizy danych chromatograficznych, które mają tendencję do ignorowania hierarchicznej struktury danych i przeprowadzania analizy na poziomie analitu.

Zdolność modelu wielopoziomowego do łączenia informacji pozwala uzyskać stabilne oszacowania parametrów modelu, które w przeciwnym razie byłyby trudne do osiągnięcia bez wykorzystania dużych zbiorów danych. Podejście to jest odmienne od zaproponowanej w pracy Haddad (2017) metody polegającej na tworzeniu ,,lokalnego'' modelu QSRR. Algorytm tej metody wyszukuje anality z pewnych (prawdopodobnie dużych) baz danych, które są podobne do analitów, dla których potrzebne są prognozy. Podobieństwo analitów jest definiowane za pomocą danej miary podobieństwa, takiej jak podobieństwo strukturalne związków (wskaźnik podobieństwa Tanimoto), podobieństwo fizykochemiczne związków (lipofilowość), obojętność, kwasowość lub zasadowość związku. Na podstawie tego ograniczonego zestawu danych budowany jest model QSRR, który jest następnie wykorzystywany do przewidywania retencji dla pożądanej grupy analitów. Takie zlokalizowane prognozy mogą być zaszumione, zwłaszcza jeśli w zbiorze danych dostępnych jest tylko kilka podobnych analitów do prognoz.

Ważnym elementem każdego modelu bayesowskiego jest definicja rozkładów *a priori* dla wszystkich jego parametrów. Dzięki temu możliwe jest włączenie wcześniejszej wiedzy o badanym zjawisku, a także regularyzacja parametrów. Jako że stanowią one istotne założenie, mogą być kwestionowane i zmieniane w zależności od różnych stanów wiedzy o problemie. Na przykład grupy funkcyjne można podzielić na podgrupy (zjonizowane, niezjonizowane) z oddzielnymi rozkładami *a priori*. W tej pracy wykorzystano słabo informacyjne rozkłady *a priori* pozwalające uzyskać spójne i stabilne rozwiązania nawet przy nieoczekiwanych wartościach parametrów.

Różne oddziaływania zachodzące w kolumnie chromatograficznej charakteryzowane są zwykle poprzez starannie zaprojektowany zestaw eksperymentów z wykorzystaniem analitów o różnych własnościach. Pozwala to na szczegółowy opis fizyczny układów chromatograficznych. Można również scharakteryzować te interakcje za pomocą danych dotyczących czasu retencji, zebranych dla stosunkowo dużej i heterogennej grupy związków poddanych analizie w szerokim zakresie warunków. Mimo to otrzymane w ten sposób parametry populacyjne i indywidualne należy traktować jako parametry „makro" (parametry opisujące ogólne zachowanie analitów w kolumnie). Takie parametry, jak np. $\log k$, maskują fizyczną rzeczywistość różnych interakcji w kolumnie chromatograficznej, uwzględniając kilka różnych efektów naraz [Gritti i Guiochon (2005b)]. Mimo to mogą być one interpretowane w ramach przyjętego modelu i wykorzystywane do predykcji.

Niezwykle cenne jest również zidentyfikowanie parametrów chromatograficznych, które są w przybliżeniu niezależne od fazy stacjonarnej, gdyż ich rozkłady mogą posłużyć jako rozkłady *a priori* do predykcji danych dla innych faz stacjonarnych (kolumn chromatograficznych). Parametrami spełniającymi to kryterium są pK_a i α (odzwierciedlają właściwości kwasowo-zasadowe związków w rozpuszczalniku). Dodatkowo S_1 i dS_1 wydają się opisywać właściwości fazy ruchomej. S_1 reprezentuje różnicę między logarytmami współczynników retencji w czystej wodzie i w 100% metanolu, a dS_1 między logarytmami współczynników retencji w 100% metanolu i w 100% acetonitrylu. Parametrami silnie zależnymi od charakterystyki fazy stacjonarnej są natomiast $\log k_w$ i $d\log k_w$ dla form neutralnych, a także parametry opisujące wpływ pH na retencję kationów i anionów (apH) czy wpływ podstawników na logarytm współczynnika retencji w czystej wodzie dla form neutralnych (π_{logk_w}).

Problem analogiczny do znalezienia warunków pozwalających uzyskać pożądaną separację spotyka się w dziedzinie farmakokinetyki populacyjnej. Przed uzyskaniem jakiegokolwiek pomiaru stężenia dla danego pacjenta, jedynym sposobem przewidzenia jego profilu farmakokinetycznego i wybrania dla niego odpowiedniej dawki jest porównanie go do innych podobnych pacjentów (z podobnymi współzmiennymi). Gdy dostępne są dane eksperymentalne, tę dodatkową wiedzę można włączyć do przewidywania profilu stężenia i dawki specyficznej dla pacjenta. Ponieważ istnieje pewna analogia między znajdowaniem dawki a poszukiwaniem pożądanego rozdzielenia w chromatografii, można zastosować podobne narzędzia. I tak chcąc przewidzieć retencję analitu w oparciu o ograniczony zastaw danych wstępnych, zasadnym wydaje się wykorzystanie podobieństwa analitu do innych analitów, które były analizowane wcześniej. Podobieństwo to można opisać, wykorzystując model hierarchiczny z deskryptorami, takimi jak liczba grup funkcyjnych, $\log P$ czy pK_a, wpływającymi na parametry opisujące retencję. Taki model umożliwia również uwzględnienie dodatkowych informacji w prognozach i podejmowaniu decyzji za każdym razem, gdy dostępne są nowe dane eksperymentalne.

Równania QSRR wyjaśniają pewną część zmienności między analitami (parametry na poziomie populacji). Jednak dokładność prognozy opartej tylko na ich podstawie jest raczej ograniczona. Aczkolwiek nawet i w tym przypadku można przewidzieć czas retencji/współczynnik retencji oraz wywnioskować wiedzę na temat prawdopodobnego chromatogramu. Stąd informacje o strukturze analitu są pomocne i jeśli to możliwe, powinny być uwzględniane przy podejmowaniu decyzji. Oczekiwane czasy retencji, biorąc pod uwagę różne źródła informacji (parametry na poziomie populacji, $\log P$, pK_a, grupy funkcyjne, masa cząsteczkowa czy wszelkie pomiary), można łatwo zwizualizować, dając w ten sposób analitykom narzędzie do podejmowania decyzji w odniesieniu do dalszych kroków w toku rozwoju danej metody. Po pierwsze mogą oni zatrzymać rozwój metody, jeśli pożądana separacja jest nieprawdopodobna. Mają również możliwość przeprowadzenia większej liczby eksperymentów, jeśli aktualne informacje są niepełne lub niewystarczające. Mogą też podjąć decyzję o zakończeniu badań, stwierdziwszy, że rozwijana metoda nie da oczekiwanych rezultatów (tzn. jest mało prawdopodobne, aby wykonanie większej liczby eksperymentów zapewniło lepszą separację). Chromatogramy niepewności można również wykorzystać do ilościowego określenia prawdopodobieństwa pomyślnego rozdzielenia lub do obliczenia oczekiwanej użyteczności następnego eksperymentu.

Badane w rozprawie modele powinny dobrze uogólniać się na inne anality, jeśli ich masa cząsteczkowa jest mniejsza niż 600 Da. Do tworzenia modeli wykorzystano duże zestawy danych dotyczących związków chemicznych, jednak nadal istniały war-

tości predyktorów, które nie wystąpiły w żadnym analicie lub występowały bardzo rzadko. Mimo tego ograniczenia zaproponowane modele pozwalają na oszacowanie efektów dla takich wartości, jednakże obarczone są one dużą niepewnością. Natomiast dokładniej oszacowano efekty dla tych wartości predyktorów, które były powszechne w analizowanych danych. Niezależnie od sytuacji modele mogą przewidywać parametry chromatograficzne i retencję analitów oraz mogą służyć jako szablon do rozwiązywania bardziej złożonych problemów spotykanych w chromatografii i dziedzinach pokrewnych. Takie podejście pozwala na ,,zindywidualizowaną'' predykcję czasu retencji.

Modele wielopoziomowe to nowa koncepcja w dziedzinie chromatografii, a ostatnie postępy w technikach obliczeniowych pozwalają na łatwe wdrożenie tych metod w praktyce. Przedstawione modele można uczynić bardziej ogólnymi i użytecznymi w codziennej praktyce np. poprzez uwzględnienie szerokości piku, innych deskryptorów, zmienności międzylaboratoryjnej czy zwiększeniu zakresu kolumn. Aby to jednak osiągnąć, konieczna byłaby większa współpraca w zakresie gromadzenia danych chromatograficznych.

Modele te mogą być wykorzystane również do odpowiedzi na pewne pytania związane z ustalaniem warunków chromatograficznych, jak np. przy jakiej zawartości modyfikatora organicznego logarytm współczynnika retencji danego związku będzie równy 1. Jest to możliwe poprzez wyznaczenie funkcji gęstości dla uzyskanych na podstawie symulacji zawartości modyfikatora organicznego, które wygenerowano z modelu wzbogaconego o wszystkie dostępne eksperymenty. Analizę tych modeli można rozszerzyć również o funkcję użyteczności, której maksimum wyznaczyłoby szansę uzyskania najlepszego chromatogramu. Funkcja ta zdefiniowana może być w oparciu o najniższy czas retencji, najwyższy czas retencji oraz różnicę czasów retencji dla krytycznej pary analitów. W przypadku zastosowanej w pracach metodyki możliwe jest podejście sekwencyjne. Uwzględnienie informacji z kolejnych eksperymentów pozwoli na dokładniejsze zbadanie przestrzeni wszystkich możliwych eksperymentów i dobór najodpowiedniejszego z nich lub określenie, że rozdzielenie nie jest możliwe.

Modelowanie wielopoziomowe wykorzystujące podejście bayesowskie wydaje się być narzędziem, które zapewnia ujednolicony schemat analizy dużych baz danych chromatograficznych.

Rozdział 6
Wnioski

Zastosowane w rozprawie modele wielopoziomowe pozwoliły w naturalny sposób uwzględnić hierarchiczną strukturę danych oraz przewidzieć retencję dla analitów nowych (jeszcze nieprzeanalizowanych) lub takich, dla których dostępna była tylko ograniczona liczba pomiarów. Co więcej, wykorzystane wnioskowanie bayesowskie umożliwiło włączyć wcześniejszą wiedzę i wykorzystać znaną teorię chromatograficzną do prognoz. Dzięki temu podejściu było możliwe obliczenie niepewności przewidywań oraz wizualizacja ich np. za pomocą chromatogramu niepewności (tj. gęstości rozkładu *a posteriori* współczynników retencji oczekiwanych dla każdego analitu w danych warunkach chromatograficznych, biorąc pod uwagę aktualną wiedzę na temat retencji). Taka wizualizacja przewidywań została omówiona pod kątem przydatności w podejmowaniu decyzji. Modele te mogą stanowić podstawę kolejnych, uwzględniających więcej zmiennych, jak np. większy zakres kolumn czy zmienność międzylaboratoryjną.

Przedstawione modele są interpretowalne i zapewniają zwięzłe podsumowanie złożonych danych. To podejście stanowi interesującą alternatywę dla różnych procedur chemometrycznych i metod uczenia maszynowego.

Bibliografia

1. Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley interdisciplinary reviews: computational statistics*, 2(1):97–106.

2. Biecek, P. (2013). *Analiza danych z programem R: modele liniowe z efektami stałymi, losowymi i mieszanymi*. Wydawnictwo Naukowe PWN.

3. Bouwmeester, R., Martens, L. i Degroeve, S. (2019). Comprehensive and empirical evaluation of machine learning algorithms for small molecule lc retention time prediction. *Analytical Chemistry*, 91(5):3694–3703.

4. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

5. Brooks, S., Gelman, A., Jones, G. L. i Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

6. Chaloner, K. i Verdinelli, I. (1995). Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304.

7. Cole, S. R., Chu, H. i Greenland, S. (2014) Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology*, 179(2):252–260.

8. Cramer, R. D. (1993). Partial Least Squares (PLS): Its strengths and limitations. *Perspectives in Drug Discovery and Design*, 1(2):269–278.

9. Croux, C., Dhaene, G., and Hoorelbeke, D. (2004). Robust standard errors for robust estimators. *CES-Discussion paper series (DPS) 03.16*, strony 1–20.

10. Dolan, J. (2006). Dwell volume revisited. *Lc Gc North America*, 24(5):458–466.

11. Dong, M. W. (2006). *Modern HPLC for practicing scientists*. John Wiley and Sons.

12. Dorsey, J. G. i Dill, K. A. (1989). The molecular mechanism of retention in reversed-phase liquid chromatography. *Chemical Reviews*, 89(2):331–346.

13. Efron, B. i Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

14. Freedman, D. (2006). On the so-called "huber-sandwich estimator" and "robust standard errors". *The American Statistician*, 60:299–302.

15. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378. Nonlinear Methods and Data Mining.

16. Gelman, A. (2003). A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing. *International Statistical Review / Revue Internationale de Statistique*, 71(2):369–382.

17. Gelman, A., Carlin, J. B., Stern, H. S. i Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*, volume 45. Chapman and Hall/CRC.

18. Gelman, A., Meng, X.-L., i Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

19. Gelman, A. i Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, strony 457–472.

20. Gelman, A., Simpson, D. i Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood.

21. Genuer, R., Poggi, J.-M. i Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

22. Goldstein, H. (2011). *Multilevel statistical models*, volume 922. John Wiley & Sons.

23. Gritti, F. (2021). Perspective on the Future Approaches to Predict Retention in Liquid Chromatography. *Analytical Chemistry*, 93(14):5653–5664.

24. Gritti, F. i Guiochon, G. (2005a). Adsorption Mechanism in RPLC. Effect of the Nature of the Organic Modifier. *Analytical Chemistry*, 77(13):4257–4272.

25. Gritti, F. i Guiochon, G. (2005b). Critical contribution of nonlinear chromatography to the understanding of retention mechanism in reversed-phase liquid chromatography. *Journal of chromatography. A*, 1099(1-2):1–42.

26. Grzenda, W. (2012). *Wstęp do statystyki bayesowskiej*. Oficyna Wydawnicza Szkoła Główna Handlowa.

27. Hackenberger, B. K. (2019). Bayes or not Bayes, is this the question? *Croatian medical journal*, 60(1):50–52.

28. Haddad, P., Shellie, R., Pohl, C., Szucs, R., Wen, Y., Talebi, M., Amos, R. I., Taraji, M., Park, S. H. i Dolan, J. (2016). Retention time prediction based on molecular structure in pharmaceutical method development: a perspective. *LCGC North America*, 34(8):550–558.

29. Haddad, P. R. (2017). Seeking the holy grail–prediction of chromatographic retention based only on chemical structures. *LC-GC North America*, 35(8):499–503.

30. Hanai, T. (1991). Structure-retention correlation in liquid chromatography. *Journal of Chromatography A*, 550:313–324.

31. Hanai, T. (1999). *HPLC: a practical guide*, volume 6. Royal Society of Chemistry.

32. Hastie, T., Tibshirani, R., Friedman, J. H. i Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

33. Haykin, S. (2009). *Neural networks and learning machines, 3/E*. Pearson Education India.

34. Hoffman, M. D. i Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

35. Hong, P. i McConville, P. R. (2018). Dwell volume and extra-column volume: What are they and how do they impact method transfer. *Waters Corporation, Milford, MA, USA*.

36. Hossain, M. S., Ong, Z. C., Ismail, Z., Noroozi, S. i Khoo, S. Y. (2017). Artificial neural networks for vibration based inverse parametric identifications: A review. *Applied Soft Computing*, 52:203–219.

37. Hox, J. (2002). *Multilevel analysis techniques and applications.* Quantitative methodology series. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

38. Hox, J. J., Moerbeek, M. i Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

39. Huang, H., Wei, X. i Zhou, Y. (2022). An overview on twin support vector regression. *Neurocomputing*, 490:80–92.

40. Jarosz, M. (red.) (2006). *Nowoczesne techniki analityczne*. Oficyna Wydawnicza Politechniki Warszawskiej.

41. Kamiński, M., Kartanowicz, R. i Gazda, K. (2004). *Chromatografia cieczowa: praca zbiorowa*. CDAiMS, Gdansk.

42. Knox, J. H. i Kaliszan, R. (1985). Theory of solvent disturbance peaks and experimentaldetermination of thermodynamic dead-volume in column liquid chromatography. *Journal of Chromatography A*, 349(2):211–234.

43. Koronacki, J. i Ćwik, J. (2008). *Statystyczne systemy uczące się*. Akademicka Oficyna Wydawnicza EXIT.

44. Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, 5(10):1282–1291.

45. Kubacki, J. (2014). Zastosowanie hierarchicznej estymacji bayesowskiej w szacowaniu wartości dochodów ludności w powiatach. *Wiadomości Statystyczne*, 9(640):21–45.

46. Lindley, D. V. (1972). *Bayesian Statistics*. Society for Industrial and Applied Mathematics.

47. Lindstrom, M. J. i Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, strony 673–687.

48. Lunn, D., Jackson, C., Best, N., Thomas, A. i Spiegelhalter, D. (2013). The bugs book. *A Practical Introduction to Bayesian Analysis, Chapman Hall, London.*

49. McNaught, A. D. i Wilkinson, A. (1997). Iupac. compendium of chemical terminology, 2nd ed. (the "gold book"). Blackwell Scientific Publications, Oxford. Online version (2019-) created by S. J. Chalk.

50. Mitra, E. D. i Hlavacek, W. S. (2019). Parameter estimation and uncertainty quantification for systems biology models. *Current Opinion in Systems Biology*, 18:9–18.

51. Montesinos López, O. A., Montesinos López, A. i Crossa, J. (2022). *Bayesian Genomic Linear Regression*, strony 171–208. Springer International Publishing, Cham.

52. Neue, U. D., Phoebe, C. H., Tran, K., Cheng, Y.-F. i Lu, Z. (2001). Dependence of reversed-phase retention of ionizable analytes on pH, concentration of organic solvent and silanol activity. *Journal of Chromatography A*, 925(1):49–67.

53. Nikitas, P. i Pappa-Louisi, A. (2005). New approach to linear gradient elution used for optimisation in reversed-phase liquid chromatography. *Journal of Chromatography A*, 1068(2):279–287.

54. Nikitas, P., Pappa-Louisi, A. i Agrafiotou, P. (2002). Effect of the organic modifier concentration on the retention in reversed-phase liquid chromatography: I. General semi-thermodynamic treatment for adsorption and partition mechanisms. *Journal of Chromatography A*, 946(1):9–32.

55. Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley and Sons.

56. Pinheiro, J. i Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

57. Rafferty, J. L., Zhang, L., Siepmann, J. I. i Schure, M. R. (2007). Retention mechanism in reversed-phase liquid chromatography: a molecular perspective. *Analytical chemistry*, 79(17):6551–6558.

58. Ranstam, J. i Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, 105(10):1348–1348.

59. Schellinger, A. P. i Carr, P. W. (2006). Isocratic and gradient elution chromatography: A comparison in terms of speed, retention reproducibility and quantitation. *Journal of Chromatography A*, 1109(2):253–266.

60. Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K. i Ramachandran, M. (2021). A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. *Processes*, 9(11):2015.

61. Shoenmakers, P. J., Billiet, H. A. i De Galan, L. (1979). Influence of organic modifiers on the rentention behaviour in reversed-phase liquid chromatography and its consequences for gradient elution. *Journal of Chromatography A*, 185:179–195.

62. Skoog, D. A., Holler, F. J. i Crouch, S. R. (2017). *Principles of instrumental analysis*. Cengage learning.

63. Snyder, L. R. i Dolan, J. W. (2006). Theory and Derivations. In *High-Performance Gradient Elution*, strony 370–413. John Wiley and Sons, Inc., Hoboken, NJ, USA.

64. Snyder, L. R., Dolan, J. W. i Lommen, D. C. (1989). Drylab® computer simulation for high-performance liquid chromatographic method development: I. Isocratic elution. *Journal of Chromatography A*, 485:65–89.

65. Snyder, L. R., Kirkland, J. J. i Dolan, J. W. (2009). *Introduction to Modern Liquid Chromatography*. John Wiley and Sons, Inc., Hoboken, NJ, USA.

66. Snyder, L. R., Kirkland, J. J. i Glajch, J. L. (1997). *Practical HPLC Method Development*. Wiley.

67. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

68. Trzęsiok, J. (2014). Porównanie zdolności predykcyjnych modelu regresji grzbietowej z wybranymi nieparametrycznymi modelami regresji. *Studia Ekonomiczne*, (191):65–74.

69. Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231.

70. Wang, R. (2012). Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800–807.

71. Wold, S., Sjöström, M. i Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130.

72. Zhang, F. i O'Donnell, L. J. (2020). Chapter 7 - support vector regression. W Mechelli, A. i Vieira, S. (red.), *Machine Learning*, strony 123–140. Academic Press.

73. Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., Goyal, H., et al. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of translational medicine*, 6(11).

74. Zou, J., Han, Y. i So, S.-S. (2009). *Overview of Artificial Neural Networks*, strony 14–22. Humana Press, Totowa, NJ.

75. Žuvela, P., Skoczylas, M., Jay Liu, J., Bączek, T., Kaliszan, R., Wong, M. W. i Buszewski, B. (2019). Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chemical Reviews*, 119(6):3674–3729.

Publikacje

Publikacja A

Link do suplementu: https://pubs.acs.org/doi/suppl/10.1021/acs.analchem.0c05227/suppl_file/ac0c05227_si_001.pdf

analytical chemistry

pubs.acs.org/ac Article

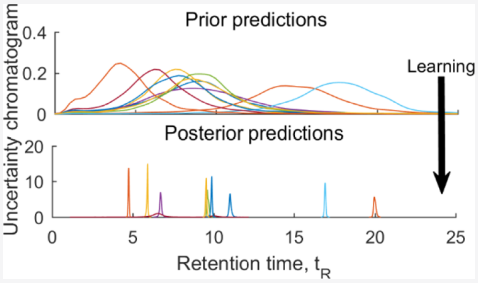
Application of Bayesian Multilevel Modeling in the Quantitative Structure−Retention Relationship Studies of Heterogeneous Compounds

Paweł Wiczling,* Agnieszka Kamedulska, and Łukasz Kubik

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

ACCESS | Metrics & More | Article Recommendations | 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

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

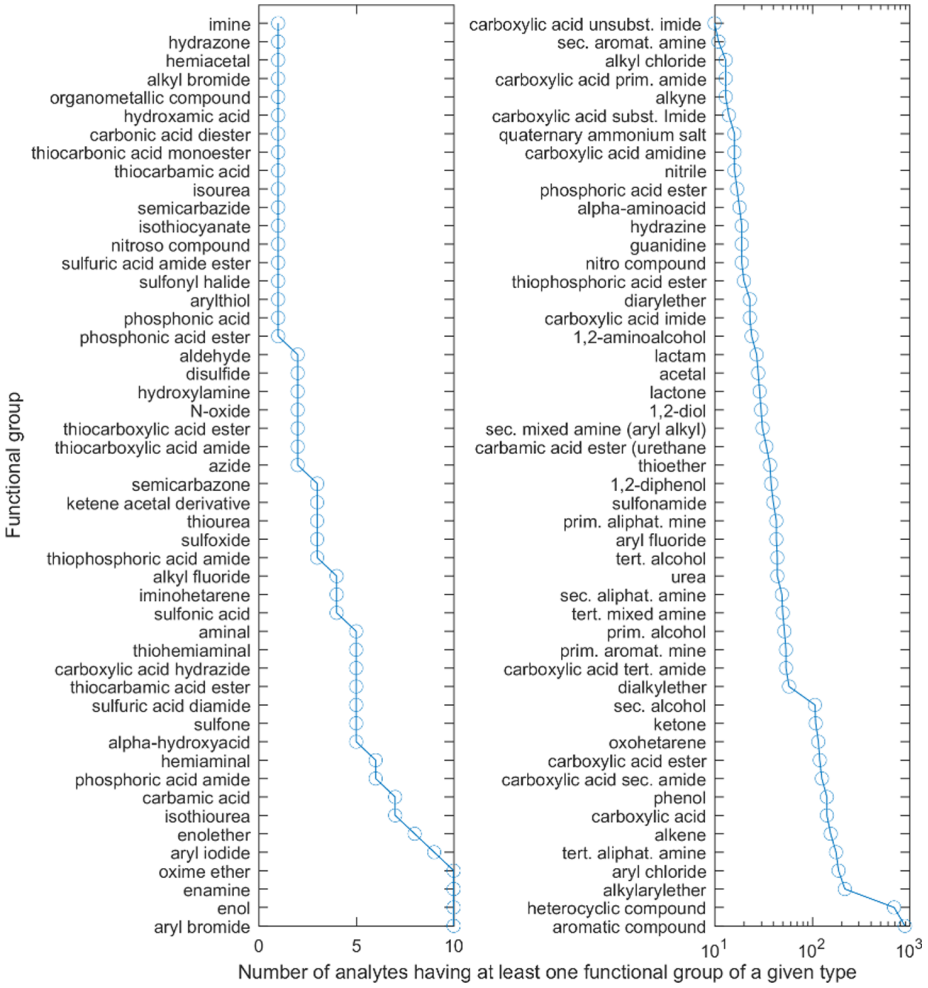


Figure 1. Functional groups identified by Checkmol. Figures show the number of analytes having at least one functional group of a given type.

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

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

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 $\log k_{w,i}$, $S_{1,i}$, $S_{2,i}$ 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

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

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 ($\sim$) 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 ($\log k_{w,i}$, $\log k_{a,i}$, 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 ($\pi_{\log S_2}$, $\pi_{\log k_w}$, and $\pi_{\mathrm{dlog}k}$ equal to the difference between $\pi_{\log k_w}$ and $\pi_{\log k_a}$)

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

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

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

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, $\pi_{\log k_w}$ 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

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

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 (5^{th}–95^{th} 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

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

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,

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

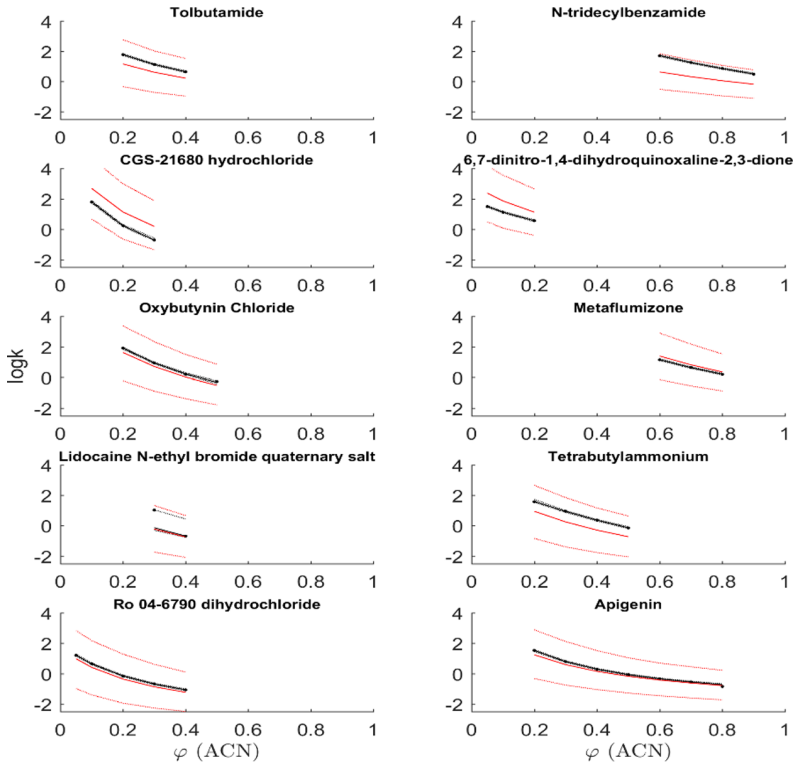


Figure 3. Individual and population predictions represented as posterior medians (lines) and 5th−95th percentiles (dotted lines) for a random set of 10 analytes. Observed retention factors are shown as dots. Black corresponds to future observations on the same analyte, and red corresponds to future observations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

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

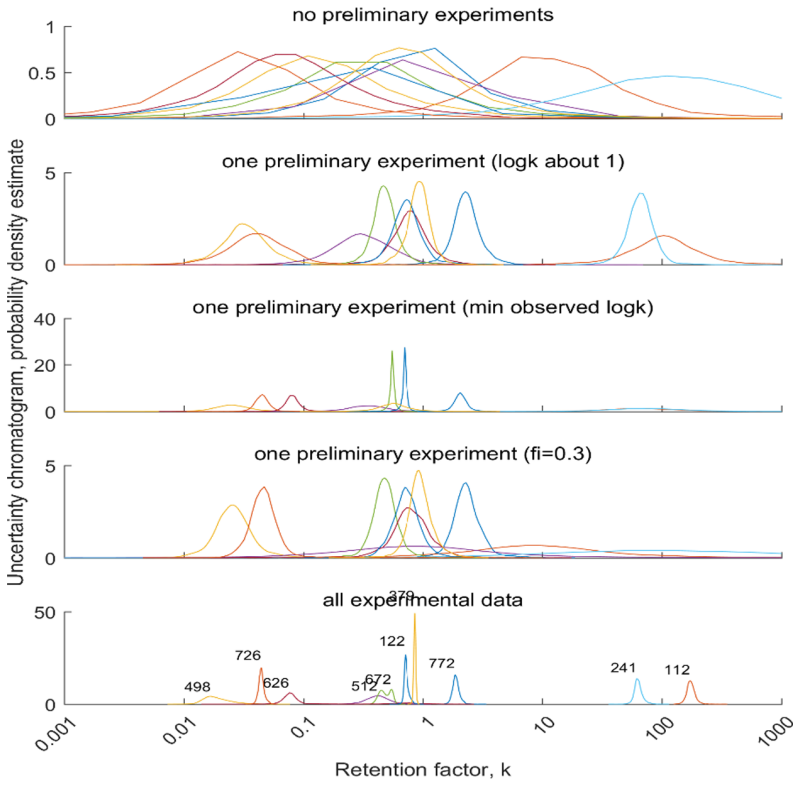


Figure 4. Uncertainty chromatograms summarizing predictions for a selection of isocratic conditions. Each peak represents the range of analyte retention factors along with uncertainty, as predicted by the proposed model conditional on different preliminary data. Colors correspond to different analytes 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).

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

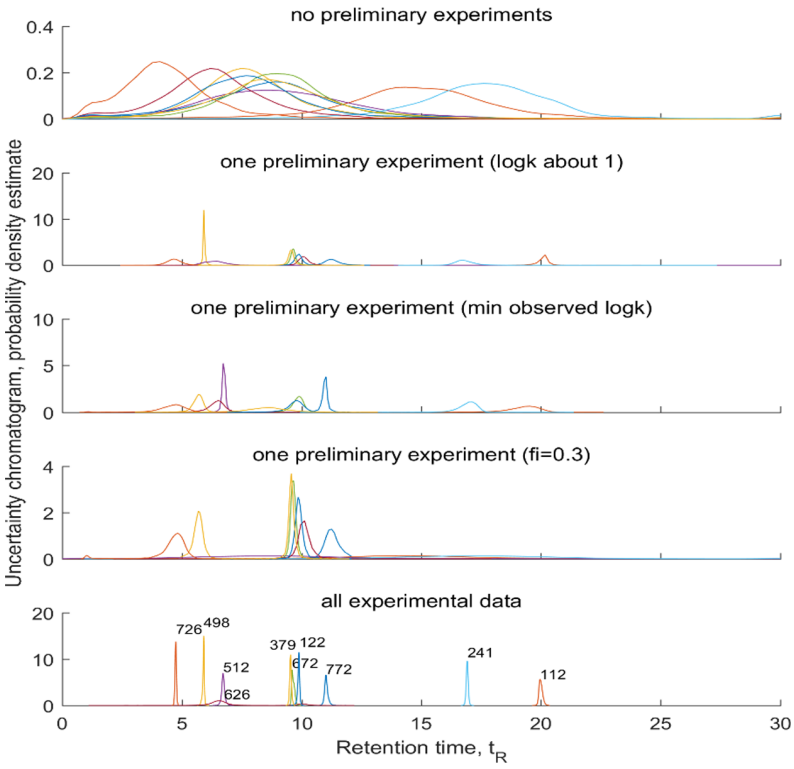


Figure 5. Uncertainty chromatograms summarizing prediction for a selection of gradient conditions. Each peak represents the range of analyte retention factors along with uncertainty, as predicted by the proposed model conditional on different preliminary data. Colors correspond to different analytes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

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

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: $\log P_{R-X} = \log P_{R-H} + \pi_R$, 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
Anal. Chem. XXXX, XXX, XXX−XXX

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

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.

https://doi.org/10.1021/acs.analchem.0c05227
Anal. Chem. XXXX, XXX, XXX−XXX

Publikacja B

Link do suplementu: https://static-content.springer.com/esm/art%3A10.1007%2Fs00216-022-03968-x/MediaObjects/216_2022_3968_MOESM1_ESM.pdf

Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-022-03968-x

RESEARCH PAPER

Statistical analysis of isocratic chromatographic data using Bayesian modeling

Agnieszka Kamedulska[1] · Łukasz Kubik[1] · Paweł Wiczling[1]

Received: 16 September 2021 / Revised: 28 January 2022 / Accepted: 8 February 2022

Abstract

Chromatographic retention times are usually modeled considering only one analyte at a time. However, it has certain limitations as no information is shared between the analytes, and consequently the model predictions poorly generalize to out-of-sample analytes. In this work, a publicly available dataset was used to illustrate the benefits of pooling the individual data and analyzing them simultaneously utilizing Bayesian hierarchical approach. Statistical analysis was carried out using the Stan program coupled with R, which enables full Bayesian inference with Markov chain Monte Carlo sampling. This methodology allows (i) incorporating prior knowledge about the likely values of model parameters, (ii) considering the between-analyte variability and the correlation between the model parameters, (iii) explaining the between-analyte variability by available predictors, and (iv) sharing information across the analytes. The latter is especially valuable when only limited information is available in the data about certain model parameters. The results are obtained in the form of posterior probability distribution, which quantifies uncertainty about the model parameters and predictions. Posterior probability is also directly relevant for decision-making. In this work, we used the Neue model to describe the relationship between retention factor and acetonitrile content in the mobile phase for 1026 analytes. The model was parametrized in terms of retention factor in 100% water, retention factor in 100% acetonitrile, and curvature coefficient, and considered log P and pK_a as predictors. From this analysis, we discovered that the analytes formed two clusters with different retention depending on the degree of analyte dissociation. The final model turned out to be well calibrated with the data. It gives insight into the behavior of analytes in the chromatographic column and can be used to make predictions for a structurally diverse set of analytes if their log P and pK_a values are known.

Keywords Retention modeling · Multilevel model · Bayesian inference · Method development

Introduction

The reversed-phase high-performance liquid chromatography (RP-HPLC) is one of the most versatile and popular analytical techniques [1]. The successful use of this method is facilitated by the understanding of how separation is influenced by various chromatographic conditions (i.e., pH, content of organic modifier, and type of stationary phase) for all analytes present in the analyzed mixture [1]. A quantitative approach of method development, utilizing mathematical models, would be an interesting alternative to the usual qualitative approach if it was proven useful for diverse problems.

Prediction of retention times is difficult due to the complexity and multiplicity of various interactions that occur during separation in an RP-HPLC column. The most important of these are the interactions between the analytes, between the solvent and the analytes, and the interactions of the stationary phase constituents with each other, as well as with the molecules of analytes and solvents [2]. The degree of these interactions depends on the chromatographic conditions and properties of analytes. Despite these complexities, some aspects of chromatographic retention are similar for structurally diverse analytes. For example, the relationship between the analyte retention factor and the chromatographic conditions can be described by the same model (i.e., the Neue model) for most analytes [2, 3]. This implies the presence of a common mechanism governing the retention of these analytes. It is also possible to relate

✉ Paweł Wiczling
wiczling@gumed.edu.pl

[1] Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, al. Gen. J. Hallera 107, Gdańsk, 80-416, Poland

Published online: 28 March 2022

the model parameters (retention factor in 100% water, retention factor in 100% acetonitrile, curvature coefficient) and the physicochemical properties of the analytes, such as lipophilicity (log P) and dissociation constant (pK_a). This implies that similar analytes (concerning these properties) will have similar retention. The retention likeness of analytes can be quantitated using hierarchical models, also known as multilevel models or mixed-effect models [4–7].

Various chemometrics/machine learning techniques (Rank Support Vector Machine, Support Vector Regression, Partial Least Squares, Kernel-Based PLS, Least Absolute Shrinkage and Selection Operator, Artificial Neural Networks, Random Forest) can be used to predict retention times [8–10]. In this work, we decided to use Bayesian hierarchical models. Hierarchical models are statistical models containing two types of parameters: population-level parameters (fixed effect) and individual-level parameters (random effect). The fixed effects are the parameters of the population which are immutable for each data collected from the population. In contrast, the random effects are those parameters whose values differ for individual representatives of the population (i.e., each analyte). As for the regression models, their purpose is to describe the response variable as a function of predictor variables. However, the advantage of multilevel models is the recognition of similarity between the analytes which provides more information about the retention of individual analytes, due to the fact that the missing information can be borrowed from other similar analytes. Classical regression models, on the other hand, completely ignore the similarity and treat each analyte separately, which induces the need to provide more data in order to obtain results similar to those from multilevel models [11]. The application of Bayesian inference to multilevel models allows using the previous knowledge about the phenomenon studied and/or regularizing the inferences through prior design to yield smoother, more stable inferences. Basically, uncertain variables can be assigned a probability distribution (prior information). This prior knowledge is often available in the literature. Including this additional information to the model may improve the accuracy and reliability of the estimated parameters [12].

Chromatographic data are usually modeled considering only one analyte at a time. However, it has certain limitations as no information is shared between the analytes and, consequently, the model predictions poorly generalize to out-of-sample analytes. In this work, a publicly available dataset was used to illustrate the benefits of pooling the individual data and analyzing them simultaneously utilizing Bayesian hierarchical approach. Statistical analysis was carried out using Stan program coupled with R, which enables full Bayesian inference with Markov chain Monte Carlo (MCMC) sampling. This methodology allows (i) incorporating prior knowledge about the likely values of the model parameters, (ii) considering the between-analyte variability and the correlation between the model parameters, (iii) explaining the between-analyte variability by available predictors (log P and pK_a), and (iv) sharing information across the analytes. The usefulness of the model in predictions was illustrated in situations of a limited amount of experimental data (i.e., none, one or two preliminary experiments).

Methods

Data

In this work, we used a publicly available dataset at www.retentionprediction.org/hplc/database/ that comprises the measurements of RP-HPLC retention times collected for 1026 analytes. The retention times were measured under isocratic conditions on Eclipse Plus C18 (Agilent) stationary phase with 3.5 µm particles. The experiments were conducted using a mixture of two solvents: solvent A, which was made of 0.1% formic acid in water, and solvent B, which was made of 0.1% formic acid in acetonitrile. The column temperature was set at 35 °C. The pH value of the mobile phase was verified experimentally for the purpose of this work. It equaled 2.66 with a standard deviation of 0.19 for the range of acetonitrile contents from 5 to 95% (Supporting Figure S1). The data were collected by Boswell et al. and were used to create a method to predict retention time by Back-Calculating the Gradient [13, 14]. The values of lipophilicity log P, molecular mass MM, and pK_a were added to the dataset. They were calculated using ACD/Labs program (www.acdlabs.com) based on the provided structures of analytes. The log P value of the analytes ranged from -5 to 8.75, MM ranged from 73.09 to 656.8 and pK_a ranged from -21.40 to 19.14. For one analyte (diphenyleneiodonium chloride), ACD/Labs program could not calculate the log P value, and therefore, this value was treated as missing, which was modeled assuming that log P of that analyte comes from the normal distribution with location 2.56 and scale 1.92, which are the mean and standard deviation of log P data for other analytes. The range of log P, MM, and pK_a defines the applicability domain of the proposed models. The raw data used in this work are shown in the Supporting Information and also presented graphically in Supporting Figure S2.

Implementation

All the models described in this work were implemented in Stan program. Stan creates representative samples of model parameters (including predictions) in the form of posterior distribution. To generate Monte Carlo steps, Stan uses the Hamiltonian Monte Carlo (HMC) algorithm and its adaptive

variant, the no-U-turn sampler. HMC avoids sensitivity to correlated parameters and random-walk behavior, which is a problem in many MCMC methods. This algorithm takes a series of steps informed by the first-order gradient information. It enables converging to high-dimensional target distributions much faster than simpler methods, such as random-walk Metropolis or Gibbs sampling [15–17].

In this work, we used RStan (version 2.19.3), which is the R interface to Stan [18]. Each model was fitted with four chains iterated at least 2000 times and warm-up equal to at least 1000.

Stan codes were inspired by the work of Margossian and Gillespie [19].

Model development procedure

The purpose of this work was to propose a model that can simultaneously describe the isocratic retention times collected for 1026 analytes. Initially, we used a model with partial pooling and a common distribution for analyte-specific parameters as described in our previous work [20]. This model assumes a nonlinear relationship between the logarithm of the retention coefficient $\log k$ and the organic modifier content φ (Neue et al. [21] equation):

$$\log k = \log k_w - \frac{S_1 \cdot \varphi}{1 + S_2 \cdot \varphi},$$

where $\log k_w$ stands for the logarithm of the retention factor corresponding to neat water as the eluent, and S_1 and S_2 are constants describing the steepness of the relationship between the solvent composition and the logarithm of the retention factor. We also assumed that:

$$S_1 = (\log k_w - \log k_a) \cdot (1 + 10^{\log S_2}),$$

where $\log k_a$ denotes the logarithm of the retention factor in 100% acetonitrile and $\log S_2$ denotes the logarithm of the curvature coefficient. The adopted hierarchical model has the following structure:

$$\log k_{Obs_{i,j}} \sim \mathcal{N}(f(R_i, \varphi_{i,j}), \sigma),$$
$$R_i \sim \mathrm{MST}(\nu, \theta_R + \beta \cdot log P_i, \Omega),$$

where $R_i = (\log k_{w_i}, \log k_{a_i}, \log S_{2_i})$ is a vector of analyte-specific chromatographic parameters, f is a function corresponding to the right side of the Neue equation, MST is a multivariate Student's t-distribution, θ_R is a vector of the expected values of R_i for the analyte with $\log P_i = 0$, and β is a vector of slopes between $\log P_i$ and R_i. In turn, σ is a scale (here standard deviation) of residuals and Ω is the scale matrix for individual-level parameters related to the unexplained between-analyte variability. The MST distribution with ν was used to ensure flexibility and robustness for analytes with unusually high or low values of R_i.

At this stage, we used weakly informative priors as described in our previous work [20]. This preliminary model showed that the distribution of $\eta_{\log k_{a,i}} = \log k_{a,i} - (\theta_{\log k_a} + \beta_{\log k_a} \cdot \log P_i)$, within-analyte residuals, is bimodal (see Supporting Figure S3). It indicates that analytes with a similar $\log P$ are grouped into two clusters.

To explain this phenomenon, the individual chromatographic parameters were assumed to follow a distribution composed of a mixture of two different MST distributions:

$$p_{R_i}\left(\begin{bmatrix}\log k_{w,i}\\ \log k_{a,i}\\ \log S_{2,i}\end{bmatrix} \middle| \theta, \Omega, \beta, \log P\right)$$
$$= \lambda \cdot \mathrm{MST}\left(7, \begin{bmatrix}\theta_{\log k_{w_1}} + \beta_{\log k_{w_1}} \cdot \log P_i\\ \theta_{\log k_{a_1}} + \beta_{\log k_{a_1}} \cdot \log P_i\\ \theta_{\log S_{2_1}} + \beta_{\log S_{2_1}} \cdot \log P_i\end{bmatrix}, \Omega_1\right) +$$
$$+ (1-\lambda) \cdot \mathrm{MST}\left(7, \begin{bmatrix}\theta_{\log k_{w_2}} + \beta_{\log k_{w_2}} \cdot \log P_i\\ \theta_{\log k_{a_2}} + \beta_{\log k_{a_2}} \cdot \log P_i\\ \theta_{\log S_{2_2}} + \beta_{\log S_{2_2}} \cdot \log P_i\end{bmatrix}, \Omega_2,\right),$$

where λ and $1-\lambda$ express the fraction of analytes belonging to the first and second cluster. Betas denote the regression coefficients for the relationship between the analyte-specific model parameters and $\log P_i$ values. Due to the appearance of new parameters, we reviewed prior distributions. The clustering of analytes seems to be caused by the presence of analytes in a neutral or dissociated form at the pH of the mobile phase used in this set of experiments. To obtain an approximate range of model parameters, we fitted the data of each analyte separately, and then combined the individual parameter estimates (the two-stage approach). Analyte-specific chromatographic parameters were estimated for each analyte using the Neue model and least-square procedure assuming $S_2 = 2$ for all analytes. This assumption was necessary to obtain stable estimates for all analytes. Then, a linear regression procedure was used to determine the parameters of a regression line between $\log k_{w,i}$ and $\log P_i$ as well as between $\log k_{a,i}$ and $\log P_i$ separately for neutral and ionized forms of analytes (illustrated in Supporting Figure S4). The intercepts and slopes of these lines were selected as the mean values of priors. The standard deviations were used to set the scale of priors. The priors for $\theta_{\log S_{2_1}}$ and $\theta_{\log S_{2_2}}$ parameters were based on the literature data suggesting a typical value of 2 for acetonitrile as a mobile phase [1, 2, 22]. Further, we assumed a standard deviation of 0.2 on a logarithmic scale with the base of 10. It corresponds to values of S_2 that range from 0.9 to 4.3

(5^{th}–95^{th} percentile). Then, we decomposed our prior scale matrix

$$\Omega = \begin{bmatrix} \omega_{\log k_w} & 0 & 0 \\ 0 & \omega_{\log k_a} & 0 \\ 0 & 0 & \omega_{\log S_2} \end{bmatrix} \cdot \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} \\ \rho_{2,1} & 1 & \rho_{2,3} \\ \rho_{3,1} & \rho_{3,2} & 1 \end{bmatrix} \cdot \begin{bmatrix} \omega_{\log k_w} & 0 & 0 \\ 0 & \omega_{\log k_a} & 0 \\ 0 & 0 & \omega_{\log S_2} \end{bmatrix}$$

into a scale ω and a correlation matrix ρ. Parameters ω represent the scales, which are non-negative. In this case, we applied the half-normal distribution (denoted as N_+) with scales determined during the two-stage approach. Since the scale for $\log k_a$ and $\log k_w$ parameters is about 1, we a priori assumed that the values of ω higher than 2 are unlikely for these parameters. To model ρ, we used the Lewandowski−Kurowicka−Joe distribution with parameter equal to one, which means that the density is uniform over 3 × 3 correlation matrices. Due to the lack of knowledge about clusters, we assumed that λ should come from the beta distribution with the parameters (1,1), which corresponds to a uniform distribution on the interval [0,1]. Our choice of prior for the standard deviation for the residuals was half-normal distribution (the same like for omegas). The standard deviation of 0.067 was selected based on the pooled residual standard error obtained from the two-stage analysis. This value is small due to the precise nature of chromatographic measurements.

The resulting priors are as follows:

$$\begin{aligned}
&\lambda \sim \text{Beta}(1, 1)\\
&\theta_{\log k_{w_1}} \sim \mathcal{N}(1.054, 1.136),\\
&\theta_{\log k_{a_1}} \sim \mathcal{N}(-3.437, 1.062),\\
&\theta_{\log k_{w_2}} \sim \mathcal{N}(2.053, 1.487),\\
&\theta_{\log k_{a_2}} \sim \mathcal{N}(-1.885, 1.006),\\
&\theta_{\log S_{2,1}},\ \theta_{\log S_{2_2}} \sim \mathcal{N}(\log 2, 0.2),\\
&\beta_{\log k_{w_1}},\ \beta_{\log k_{w_2}} \sim \mathcal{N}(0.7, 0.25),\\
&\beta_{\log k_{a_1}},\ \beta_{\log k_{a_2}} \sim \mathcal{N}(0.3, 0.25),\\
&\beta_{\log S_{2_1}},\ \beta_{\log S_{2_2}} \sim \mathcal{N}(0, 0.25),\\
&\omega_{\log k_{w_1}} \sim \text{N}_+(0, 1.136),\ \omega_{\log k_{a_1}} \sim \text{N}_+(0, 1.487),\\
&\omega_{\log S_{2_1}} \sim \text{N}_+(0, 0.2),\\
&\omega_{\log k_{w_2}} \sim \text{N}_+(0, 1.062),\ \omega_{\log k_{a_2}} \sim \text{N}_+(0, 1.006),\\
&\omega_{\log S_{2_2}} \sim \text{N}_+(0, 0.2),\\
&\begin{bmatrix} 1 & \rho_{1,2_1} & \rho_{1,3_1} \\ \rho_{2,1_1} & 1 & \rho_{2,3_1} \\ \rho_{3,1_1} & \rho_{3,2_1} & 1 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{1,2_2} & \rho_{1,3_2} \\ \rho_{2,1_2} & 1 & \rho_{2,3_2} \\ \rho_{3,1_2} & \rho_{3,2_2} & 1 \end{bmatrix} \sim \mathcal{LKJ}(1),\\
&\sigma \sim \text{N}_+(0, 0.067).
\end{aligned}$$

During the model, we used Student's t-distribution with a predetermined number of degrees of freedom $\nu = 7$. Therefore, we limited the number of parameters allowing for the occasional outlying values in R_i. The model was also more stable.

This model allows determining the probability of belonging to a particular cluster (calculations can be found in the Supporting Information). For graphical display, we assumed that the analytes with a probability greater than 0.5 belong to the first cluster, while the remaining ones belong to the second cluster. Figure 1 shows the distribution of random variables ($\eta_{R,i}$) describing the unexplained between-analyte variability of chromatographic parameters ($\eta_{R,i} = R_i - (\theta_R + \beta \cdot \log P_i)$). This result confirms the assumption that grouping of the studied analytes is related to the degree of analyte dissociation. The model also showed that for some analytes, the probability of belonging to the first cluster is close to 0.5. It indicates that those analytes are partially dissociated at this particular pH.

Since the cluster membership is driven by the degree of analyte dissociation, we decided to include pK_a values of analytes as an additional predictor. For a monoprotic acid, the logarithm of the ratio of the ionized form to the neutral one is given by:

$$\log\left(\frac{fr_i'}{1 - fr_i'}\right) = pH - pK_{a,i},$$

where fr_i' denotes the ratio of ionized to total concentration of a compound ($\frac{[A^-]}{[A^-]+[HA]}$) as predicted using pK_a from ACD/Labs software. Since pK_a and pH are measured with an error, it is convenient to introduce an error term at the scale of (pH and pK_a) and convert back to the adjusted fraction denoted fr_i:

$$fr_i = \frac{1}{1 + 10^{-\left(\log\left(\frac{fr_i'}{1-fr_i'}\right)+\eta_i\right)}},$$
$$\eta_i \sim t\text{-Student}(7, 0, 0.1),$$

where η_i is a random variable corresponding to the error described by Student's t-distribution with a mean of 0, scale of 0.1, and 7 degrees of freedom. A very small scale was used to indicate that the most likely pK_a values are close to the true values (accurately predicted by ACD/Labs program). Nevertheless, the presence of some outlying values was handled by Student's t-distribution. The reasoning presented here for the monoprotic analytes was extended to polyprotic analytes as shown in the Supporting Information.

This model estimates the parameters $\log k_w$, $\log k_a$, and $\log S_2$ as the following sum, taking into account the analyte-specific fr_i values:

$$\begin{bmatrix} \log k_{w,i} \\ \log k_{a,i} \\ \log S_{2,i} \end{bmatrix} = fr_i \cdot \mathrm{MST}\left(7, \begin{bmatrix} \theta_{\log k_{w_1}} + \beta_{\log k_{w_1}} \cdot \log P_i \\ \theta_{\log k_{a_1}} + \beta_{\log k_{a_1}} \cdot \log P_i \\ \theta_{\log S_{2_1}} + \beta_{\log S_{2_1}} \cdot \log P_i \end{bmatrix}, \Omega_1\right) + (1 - fr_i) \cdot \mathrm{MST}\left(7, \begin{bmatrix} \theta_{\log k_{w_2}} + \beta_{\log k_{w_2}} \cdot \log P_i \\ \theta_{\log k_{a_2}} + \beta_{\log k_{a_2}} \cdot \log P_i \\ \theta_{\log S_{2_2}} + \beta_{\log S_{2_2}} \cdot \log P_i \end{bmatrix}, \Omega_2\right).$$

It should be noted that this is a simplification. Theoretically, the retention factor should be expressed by:

$$k_i = fr_i \cdot k_{ionized} + (1 - fr_i) \cdot k_{neutral}.$$

Nevertheless, for this particular data, the formulas are practically equivalent, since almost all compounds exist either as neutral or as dissociated form. Only 14% of analytes are expected to be partially dissociated (<90% of the neutral or dissociated form present in a solution). Clearly, to fully elucidate the effects of pH on retention, more extensive data are needed, e.g., those collected for a wide range of pH values of the mobile phase.

In this model, we adopted the same priors as in the mixture model described above.

The predictive power of the model was assessed with the Watanabe–Akaike information criterion (WAIC) and root mean square error (RMSE). These measures estimate out-of-sample expectation and are conceptually similar to the deviance information criterion. A smaller WAIC and RMSE indicate a better predictive performance of the model. Because WAIC and other measures cannot assess the model performance for new analytes (they approximate leave-one-measurement-out cross-validation), a WAIC and RMSE were also calculated for population predictions (in this case, they approximate leave-one-analyte-out cross-validation). We also assessed the applicability of this model in predictions in a situation of limited experimental data. In this case, a group of 16 analytes was chosen from the original data (four of them had the worst fit parameters, four had the best, and eight had the average). This group was excluded from the analysis. The remaining analytes plus a limited number of measurements (lack of one or two preliminary experiments) from the excluded analytes were used to obtain predictions [23, 24].

Results and discussion

In this work, we present the process of building Bayesian multilevel models describing the relationship between the retention factor and the acetonitrile content for a set of isocratic measurements for 1026 analytes. We show two models with varying degrees of complexity. We started the analysis with a model including $\log P$ as a predictor. The observed grouping of analytes into two clusters was explained by the degree of dissociation determined from the pK_a values of the analytes. The models were implemented in Stan program which allows full Bayesian inference with MCMC sampling.

The relationship between the individual (analyte-specific) model parameters and $\log P$ is summarized in Fig. 2. These values can be compared to the ones obtained during the two-stage approach (Supporting Figure S4) that resembles the classical (one analyte at a time) way of analyzing this type of data. Clearly, the multilevel model leads to a "shrinkage" of extreme analyte-specific model parameters toward the prior mean (reflected by the blue line in Fig. 2).

The summary of the marginal posterior distributions of model parameters is presented in the Supporting Information (Table S1 for Model 1 and Table S2 for Model 2). Similar values of model parameters were obtained for both models. Specifically for Model 2, the value of $\log k_w$ of an analyte with $\log P = 0$ was 1.12 (CI: 0.88 to 1.36) for the first cluster (ionized form) and 2.57 (CI: 2.36 to 2.78) for the second cluster (neutral form), with CI standing for credible interval. For the $\log k_a$ parameter, a lower value of -2.95 (CI: -3.14 to -2.77) was obtained for the first cluster (ionized form) and -1.21 (CI: -1.26 to -1.16) for the second cluster (neutral form). For both parameters, the difference between the neutral and the ionized form is about one. It is in agreement with the literature findings where the typical ratio of the retention factors of the neutral and the ionized form is on the order of 10 [21]. The difference between $\log k_w$ and $\log k_a$ reflects the total free energy of transfer from water to acetonitrile. It equals 4.07 or 3.78 on average for an analyte with $\log P = 0$, depending on the cluster. This difference is higher for the analytes with larger $\log P$ values indicating the increased preferences of more lipophilic substances to acetonitrile.

We also notice a small difference in curvature coefficient between the clusters (2.9 for the first cluster and 4.1 for the second cluster). The further the value of S_2 coefficient is from zero, the less a given curve resembles a straight line. Thus, the curve of $\log k$ vs. φ relationship for the analytes from the first cluster is more flattened. The results show that there is no strong correlation between the unexplained variance corresponding to parameters $\log k_w$, $\log k_a$, and

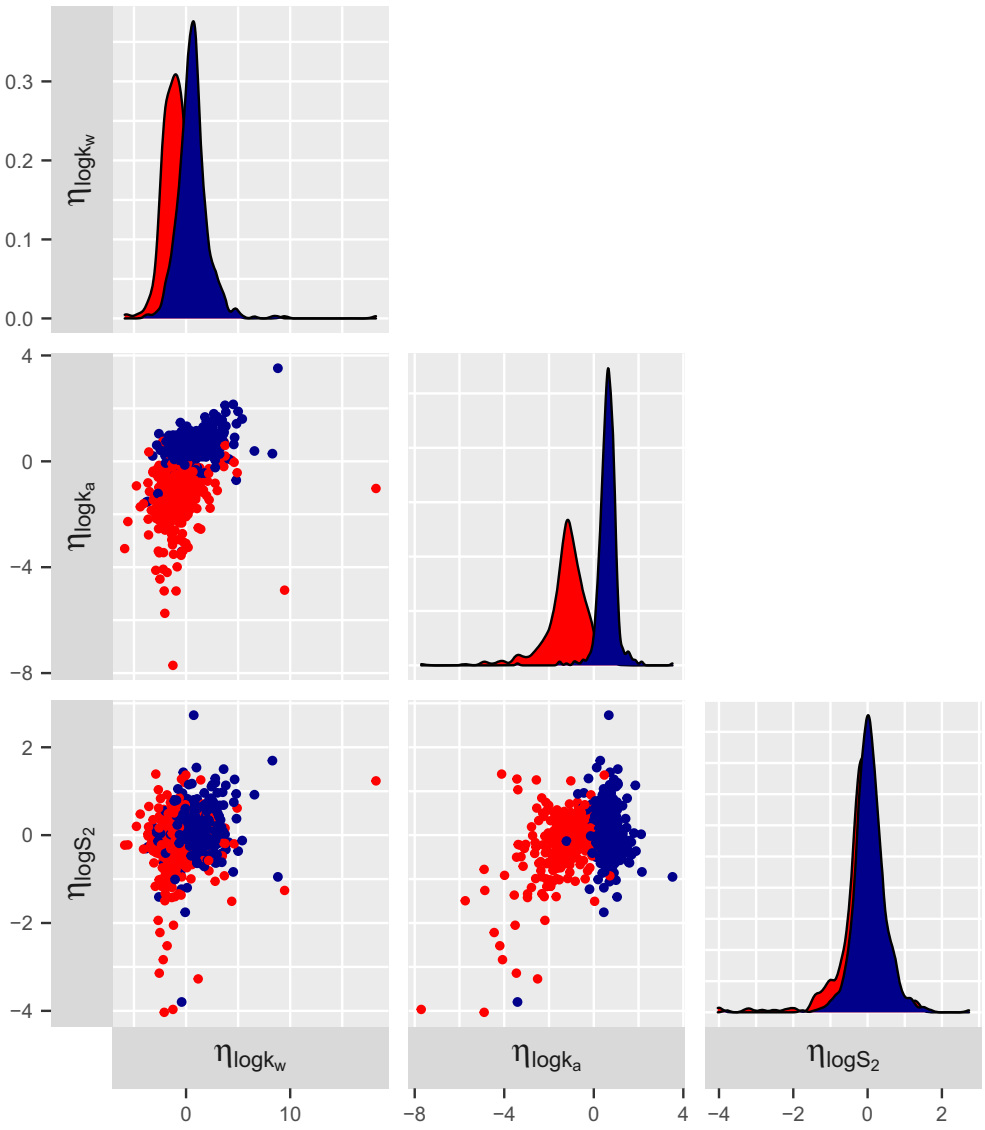


Fig. 1 Distributions of random effects. Parameters $\eta_{\log k_w}$, $\eta_{\log k_a}$, and $\eta_{\log S_2}$ divided into two clusters based on Model 1 predictions

$\log S_2$. The highest correlation was observed between the $\log k_a$ and $\log S_2$ parameters for the ionized forms of analytes.

The average $\beta_{\log k_{w_1}}$ and $\beta_{\log k_{w_2}}$ equal 0.87 (CI: 0.79 to 0.95) for the ionized form and 0.85 (CI: 0.78 to 0.92) for the neutral one. The values are close to one, which means that the values of $\log P$ are shifted from $\log k_w$ values by a constant. It is in agreement with theoretical expectations [25]. The average slopes for acetonitrile $\beta_{\log k_{a_1}}$ and $\beta_{\log k_{a_2}}$ are smaller than those for water-rich mobile phases and equal 0.25 (CI: 0.20 to 0.31) for the ionized and 0.20 (CI: 0.19 to 0.22) for the neutral form. The values of $\beta_{\log S_{2_1}}$ and $\beta_{\log S_{2_2}}$ are approximately zero (−0.01 and −0.05 for the ionized and the neutral form, respectively).

The ω parameter can be interpreted as the deviation of the particular parameter value from the expected value or between-analyte variability not explained by predictors. The largest scale was noted for the $\log k_w$ parameter, and the smallest for $\log S_2$. The scale for $\log k_w$ of 1.3 (CI: 1.19 to 1.42) and 1.5 (CI: 1.39 to 1.62) indicates that a large portion of variability is not explained by $\log P$ and pK_a values. The small values for $\log S_2$ indicate that these parameters are similar across analytes since, as expected theoretically, they reflect the influence of solvent on retention. The scale of the $\log k_a$ parameter varies greatly between the two clusters (ionized 1.01 (CI: 0.93 to 1.09), neutral 0.36 (CI: 0.33 to 0.39)). This likely indicates a more complex interaction of ionized analyte at the interface of the mobile and stationary phase in acetonitrile-rich mobile phases, i.e., the presence of secondary interactions in a form of low-energy and high-energy sites for ionized forms [26]. The low value of $\omega_{\log k_{a,2}}$ indicates that the highest accuracy of retention factor predictions can be expected for neutral forms of the analytes in acetonitrile-rich mobile phases.

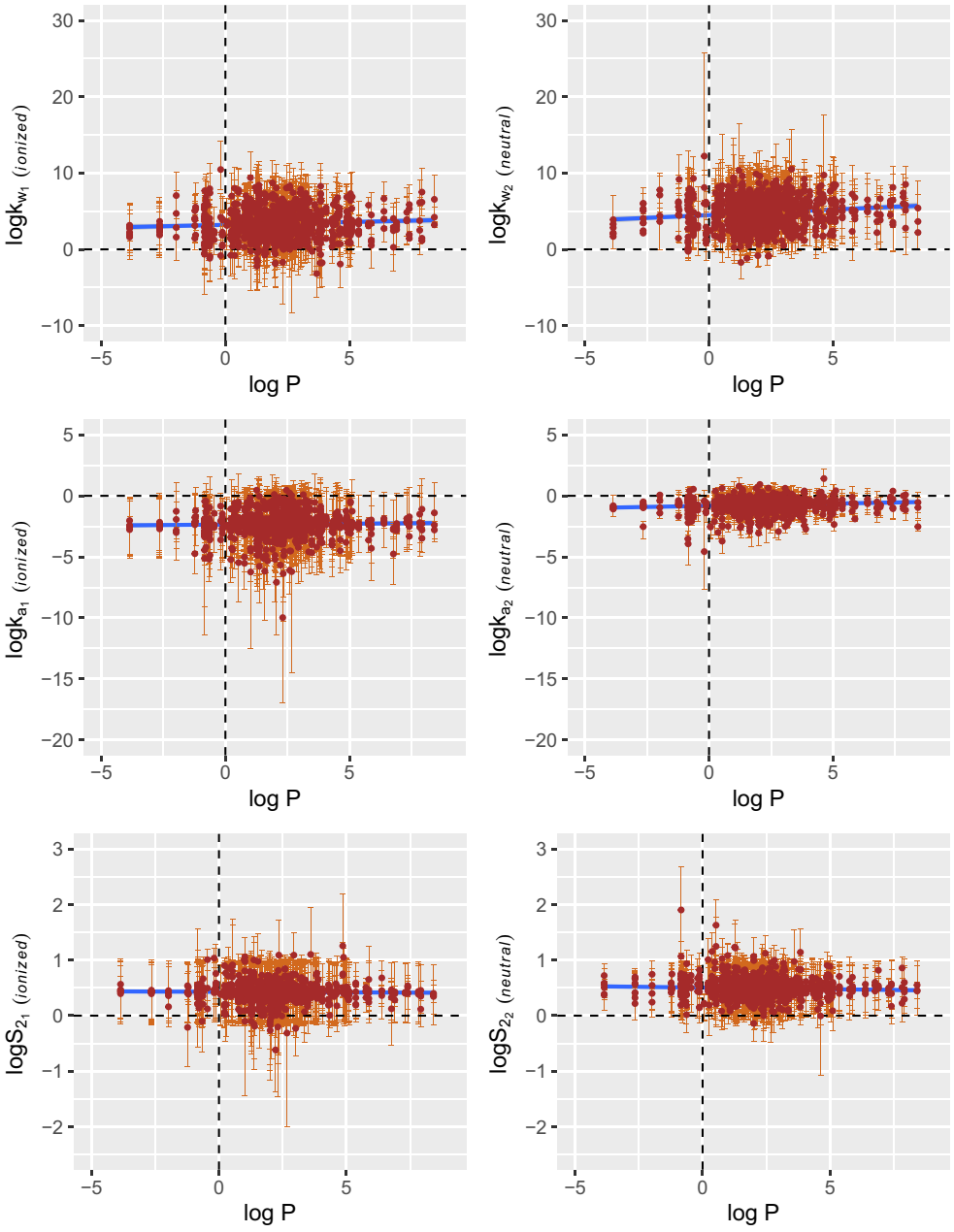


Fig. 2 Visualization of the parameters predicted by Model 2 for each analyte with the mean and standard deviation. Prediction corresponds to the future observations of the same analyte. The blue line is the line of linear regression

In both models, the missing log P value for diphenyleneiodonium chloride was a priori assumed to follow a normal distribution with a mean of 2.56 and a standard deviation of 1.92. The mean value of the missing log P estimated by Model 2 was 0.85, and the standard deviation was 1.49. This indicates that 95% of the possible log P values lie in the range between −2.13 and 3.83. It shows that chromatographic measurements provide rather approximate information about log P values. The measured log P value for diphenyleneiodonium chloride is 2.15 [27].

Despite the fact that the data and model are insufficient to fully elucidate the absorption mechanism, they provide an approximate answer for the likely retention given various sources of information (based on the predictors and population-level parameters or any number of measurements). Model 1 is useful in predicting the retention times of analytes for which log P is known, and there is no information about pK_a. The lack of knowledge of pK_a leads to less precise predictions, due to the fact that cluster membership has to be predicted from the data. On the contrary, Model 2 is useful in predicting the retention times of analytes for which log P and pK_a values are known. In this scenario, there is no need to estimate the degree of dissociation. In Figure 3, the goodness-of-fit plots are presented to assess how well the proposed models fit the data. Both plots show the relationship between the observations and model

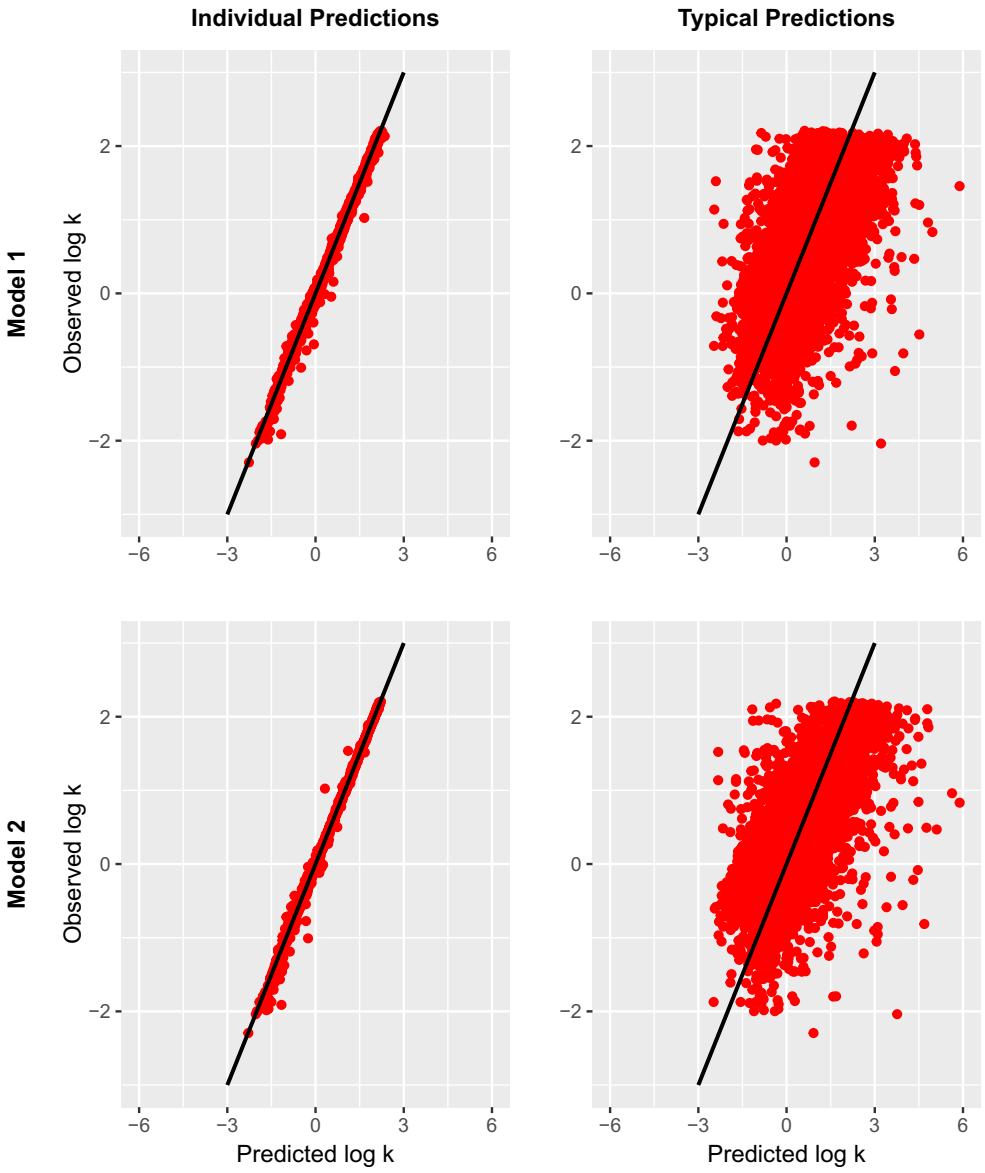


Fig. 3 The goodness-of-fit plots for Model 1 and Model 2: the observed vs. the mean individual-prediction retention factors (i.e., the a posteriori mean of a predictive distribution conditioned on the observed data from the same analyte) and the observed vs. the mean typical-prediction retention factors (i.e., the a posteriori means of predictive distributions corresponding to the future observations of a new analyte)

predictions and allow assessing the accuracy of calibration and precision of predictions. Two types of predictions were considered: population (corresponding to the future observations of a new analyte) and individual (corresponding to the future prediction of the same analyte, e.g, conditional on all observed retention time data). We note that for both models, individual predictions are highly accurate (points appear close to the line of identity) and highly precise (all points are close together). The typical predictions are less accurate. It is expected given the fact that predictions are conditional on predictors, model structure, and population-level model parameters (not on analyte-specific retention factors).

This work is an extension of the model described in the article [28]. This model used molecular weight and functional groups as predictors. In order to compare it with models presented in this paper, we mark it as Model 3. All three models gave similar goodness-of-fit plots for individual predictions (Supporting Figure S5). When it comes to population predictions, Model 3 has a slightly lower accuracy than the others (the points are more scattered around the straight line). In the case of WAIC, all three models give fairly similar results (−15534.11 for Model 1, −15239.31 for Model 2 and −15434.0 for Model 3) for individual predictions. However, the RMSE for Model 3 is higher than for the other models (Model 1: 0.052, Model 2: 0.056, Model 3: 0.308), which proves that the predictions of Model 3 are less precise. In the case of population predictions, Model 2 has lower WAIC (621324.1 vs. 474424.1 vs. 70529.0) and RMSE (1.419 vs. 1.141 vs. 1.553) values in comparison to Model 1 and 3, which indicates a better estimate of out-of-sample deviance (Supporting Table S3). The lesser precision of Model 3 is probably due to the fact that the

lipophilicity and pKa provide a more accurate information than molecular mass and functional groups (substituents).

The proposed in this work models describe the important features of chromatographic data and can be used to predict the retention factor of new analytes that are similar to those used to build the model and tested in similar conditions (i.e., on a similar column). Figure 4 shows the typical predictions, which correspond to future observations of a new analyte. The first row contains the worst predictions, and the last contains the most accurate predictions. Since predictions are based on the information about $\log P$ and pK_a and information shared from other analytes (population-level parameters), as illustrated in the first column, these are subject to considerable uncertainty. For a few analytes, the predictions do not match the observation well (these outlying analytes are handled by the MST distribution). The accuracy of predictions can be improved by collecting more data as illustrated in the second column. In these graphs, single measurement was assumed to be known. We note that the uncertainty of predictions has decreased. When we add information from the two most extreme measurements to model, we observe an even higher decrease in the uncertainty of predictions, as shown in the third column. These graphs show that the proposed model can propagate uncertainty given various number of measurements and as such can be useful for decision-making under uncertainty. The finding also supports the statement that it is rather hopeless to provide chromatographic predictions based on the knowledge of $\log P$ and pK_a which are accurate enough to be practically useful. Nevertheless, the knowledge of

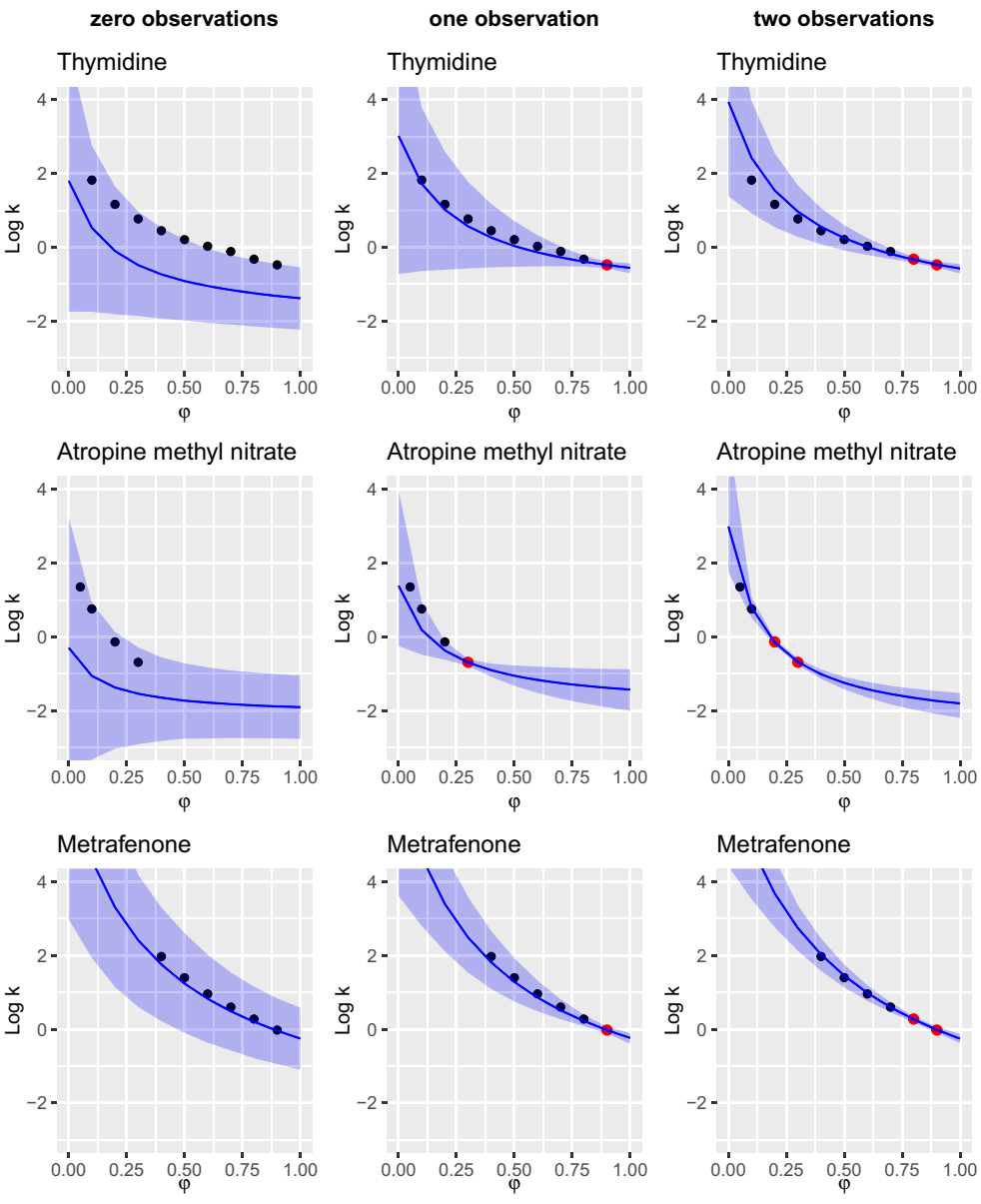


Fig. 4 Predicted (posterior median (line) and 95% credible intervals (shaded area) from Model 2 and observed logarithm retention factors for 3 representative analytes. The black dots are the experimental points and the red points are the points used in predicting the curves. Prediction corresponds to the future observations of a new analyte (i.e., posterior predictive distributions)

log P and pK_a and one (or more) scouting measurement leads to uncertainty in predictions that is small enough to make them useful. We are fully aware that the problem of predicting the RP-HPLC retention factor is far more complex than proposed in this work. A model proposed for the data obtained on a single column under isocratic conditions does not provide a directly applicable solution to usual problems encountered in every day practice. Nevertheless, we believe that developing such models for a large body of columns, laboratories, and various conditions will offer a great help in predicting analyte retention utilizing analyte structure and a limited number of preliminary experiments.

Conclusions

The dataset with isocratic measurements for 1026 analytes was analyzed using single Bayesian multilevel approach. The analysis provided the concise description of the data and several insights into the behavior of analytes in the chromatographic column. It showed the importance of taking the degree of dissociation into account when analyzing isocratic data at constant pH. It also revealed that the information provided by log P and pK_a is rather limited to make precise prediction of retention factor. In addition, it illustrated the usefulness of the Bayesian approach in calculating the uncertainty in predicted retention factors when there are limited experimental data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-022-03968-x.

Funding This study was supported by (i) the project POWR.03.02.00-00-I035/16-00 co-financed by the European Union through the European Social Fund under the Operational Programme Knowledge Education Development 2014–2020 and (ii) the National Science Centre, Poland (grant 2015/18/E/ST4/00449).

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Snyder LR, Kirkland JJ, Dolan JW. Introduction to modern liquid chromatography, 2nd ed. New York: John Wiley & Sons, Inc.; 2009.
2. 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.
3. Neue UD. Nonlinear Retention Relationships in Reversed-Phase Chromatography. Chromatographia. 2006;63(S13):S45–S53. https://doi.org/10.1365/s10337-006-0718-9, http://www.springerlink.com/index/10.1365/s10337-006-0718-9.
4. Gelman A. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. Technometrics. 2006;48(3):432–435. https://doi.org/10.1198/004017005000000661.
5. Hox J. Multilevel analysis: Techniques and applications, 2nd ed. New York: Routledge; 2010.
6. Stangl DK. Prediction and decision making using Bayesian hierarchical models. Stat Med. 1995;14(20):2173–2190.
7. Wiczling P. Analyzing chromatographic data using multilevel modeling. Anal Bioanal Chem. 2018;410(16):3905–3915. https://doi.org/10.1007/s00216-018-1061-3.
8. Haddad PR, Taraji M, Szücs R. Prediction of Analyte Retention Time in Liquid Chromatography. Anal Chem. 2021;93(1):228–256. https://doi.org/10.1021/acs.analchem.0c04190.
9. Bouwmeester R, Gabriels R, Hulstaert N, Martens L, Degroeve S. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods. 2021;18(11):1363–1369. https://doi.org/10.1038/s41592-021-01301-5.
10. Giese SH, Sinn LR, Wegner F, Rappsilber J. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. Nat Commun. 2021;12(1):3237. https://doi.org/10.1038/s41467-021-23441-0.
11. McElreath R. Statistical rethinking: a bayesian course with examples in r and stan. 2016.
12. Gelman A, Simpson D, Betancourt M. The prior can often only be understood in the context of the likelihood. Entropy. 2017;19(10):555. https://doi.org/10.4324/9781315650982.
13. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. J Chromatogr A. 2011;1218(38):6742–6749. https://doi.org/10.1016/J.CHROMA.2011.07.070, https://www.sciencedirect.com/science/article/abs/pii/S0021967311011095?via%3Dihub.
14. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. A study on retention 'projection' as a supplementary means for compound identification by liquid chromatography–mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments. J Chromatogr A. 2011;1218(38):6732–6741. https://doi.org/10.1016/J.CHROMA.2011.07.105, https://www.sciencedirect.com/science/article/abs/pii/S0021967311011447?via%3Dihub.
15. Kruschke JK. Doing bayesian data analysis: A tutorial with r, jags, and stan, 2nd ed. London: Academic Press; 2014.
16. Hoffman MD, Gelman A. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014;15(1):1593–1623.
17. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. Journal of Statistical Software, Articles. 2017;76(1):1–32. https://doi.org/10.18637/jss.v076.i01, https://www.jstatsoft.org/v076/i01.
18. Stan Development Team. RStan: the R interface to Stan. 2021. https://mc-stan.org/, R package version 2.21.3.
19. Margossian C, Gillespie B. Differential equations based models in stan. 2017. https://mc-stan.org/events/stancon2017-notebooks/stancon2017-margossian-gillespie-ode.html.
20. Kubik L, Kaliszan R, Wiczling P. Analysis of Isocratic-Chromatographic-Retention Data using Bayesian Multilevel Modeling. Anal Chem. 2018;90(22):13670–13679. https://doi.org/10.1021/acs.analchem.8b04033.
21. Neue UD, Phoebe CH, Tran K, Cheng Y-F, Lu Z. Dependence of reversed-phase retention of ionizable analytes on pH, concentration of organic solvent and silanol activity. J Chromatogr A. 2001;925(1):49–67. https://doi.org/10.1016/S0021-

9673(01)01009-3, http://www.sciencedirect.com/science/article/pii/S0021967301010093.
22. Pappa-Louisi A, Nikitas P, Balkatzopoulou P, Malliakas C. Two- and three-parameter equations for representation of retention data in reversed-phase liquid chromatography. J Chromatogr A. 2004;1033(1):29–41. https://doi.org/10.1016/J.CHROMA.2004.01.021, https://www.sciencedirect.com/science/article/pii/S0021967304000718.
23. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Stat Comput. 2014;24(6):997–1016. https://doi.org/10.1007/s11222-013-9416-2, http://link.springer.com/10.1007/s11222-013-9416-2.
24. Vehtari A, Gelman A, Gabry J. Practical bayesian model evaluation using leave-one-out cross-validation and waic. Stat Comput. 2017;27:1413–1432.
25. Hanai T. Structure—retention correlation in liquid chromatography. J Chromatogr A. 1991;550:313–324. https://doi.org/10.1016/S0021-9673(01)88547-2, http://www.sciencedirect.com/science/article/pii/S0021967301885472.
26. 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.
27. Royal Society of Chemistry. CSID:2015292. 2021. https://www.chemspider.com/Chemical-Structure.2015292.html.
28. Wiczling P, Kamedulska A, Kubik L. Application of Bayesian Multilevel Modeling in the Quantitative Structure—Retention Relationship Studies of Heterogeneous Compounds. Anal Chem. 2021;93(18):6961–6971. https://doi.org/10.1021/acs.analchem.0c05227.

Publikacja C

Link do suplementu: https://pubs.acs.org/doi/10.1021/acs.analchem.2c02034?goto=supporting-info

analytical chemistry

pubs.acs.org/ac Article

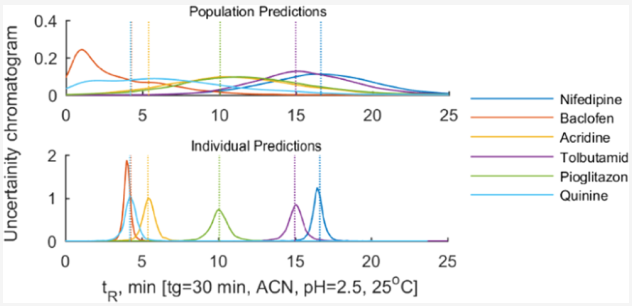
Toward the General Mechanistic Model of Liquid Chromatographic Retention

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

Cite This: https://doi.org/10.1021/acs.analchem.2c02034 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Large datasets of chromatographic retention times are relatively easy to collect. This statement is particularly true when mixtures of compounds are analyzed under a series of gradient conditions using chromatographic techniques coupled with mass spectrometry detection. Such datasets carry much information about chromatographic retention that, if extracted, can provide useful predictive information. In this work, we proposed a mechanistic model that jointly explains the relationship between pH, organic modifier type, temperature, gradient duration, and analyte retention based on liquid chromatography retention data collected for 187 small molecules. The model was built utilizing a Bayesian multilevel framework. The model assumes (i) a deterministic Neue equation that describes the relationship between retention time and analyte-specific and instrument-specific parameters, (ii) the relationship between analyte-specific descriptors (log P, pK_a, and functional groups) and analyte-specific chromatographic parameters, and (iii) stochastic components of between-analyte and residual variability. The model utilizes prior knowledge about model parameters to regularize predictions which is important as there is ample information about the retention behavior of analytes in various stationary phases in the literature. The usefulness of the proposed model in providing interpretable summaries of complex data and in decision making is discussed.



Analysis of multicomponent mixtures using liquid chromatography coupled to mass spectrometry (LC/MS) yields large datasets of retention times. Although such datasets are relatively easy to collect, they are usually heterogeneous, messy, and can contain missing records or even systematic errors, e.g., due to a lack of precise control over the experiments, instruments, and data collection process. Data analysis often requires preprocessing in the form of data cleaning and filtering. Additionally, relatively complex analysis is required to extract useful information, e.g., due to the presence of analytes with different retention characteristics and various sources of variation in the data. Either empirical (statistical) models or mechanistic (generative) models can be used to describe such datasets. Empirical models are usually built based on convenience without connections to the available chromatographic theory. On the other hand, mechanistic models describe how the observed data could have arisen from the principles and fundamentals of liquid chromatography. Such models are more appropriate for extrapolations. Regardless of the approach, the accurate prediction of the retention time in liquid chromatography is required for rapid column screening, computer-assisted method development and method transfer, and unambiguous compound identification by LC/MS analyses.[1]

Regression models are usually employed to predict retention time based on a set of scouting (preliminary) experiments and/or various predictors, such as the molecular structure or physicochemical properties of compounds.[2] There are many methods for fitting a model. However, for more complex problems involving large datasets, supervised machine learning algorithms, such as different variants of partial least squares regression (PLS),[3,4] support vector regression (SVR),[5] Bayesian ridge regression (BRR), least absolute shrinkage and selection operator regression (LASSO),[6] adaptive boosting (AB), gradient boosting (GB), random forest (RF),[7] and artificial neural network (ANN), are often employed.[1,8,9] Most of these methods are based on classical statistics. However, Bayesian methods are gaining an increasing amount of popularity, as they allow us to take into account previously acquired knowledge about the process and to quantify the uncertainty of model parameters and predictions.[10−14] Bayesian methods can also be helpful in a search for a desired separation under uncertainty.[15,16]

Received: May 10, 2022
Accepted: July 14, 2022

In this work, we aimed to propose a general mechanistic model based on LC/MS data collected for 84 controlled experiments using a mixture of 300 small-molecule analytes. The gradient experiments differed with respect to pH, type of organic modifier, gradient duration, and temperature. The data were described using the Bayesian multilevel framework.[12−14] Multilevel (hierarchical) models are statistical models that contain two types of parameters: population-level parameters and individual-level parameters. Population parameters are the same for each analyte belonging to a certain population (set) of analytes. In contrast, individual-level parameters differ for each analyte.[17] The multilevel model usually assumes that the same deterministic equation describes the relationship between retention time and analyte-specific and instrument-specific parameters, the relationship between analyte-specific descriptors ($\log P$, pK_a, and functional groups) and chromatographically specific parameters of the analyte, and various stochastic components. The benefit of such a model is that it considers the natural nesting and various heterogeneities of the data. The Bayesian method of modeling also allows the use of prior knowledge about the analyzed phenomenon and allows the building of complex models that include all of the known and relevant facts related to the particular problem one aims to solve. Such models also quantify the uncertainty of model parameters and predictions, which is particularly useful for solving problems with limited amounts of experimental data.

This paper is organized as follows: In the following section, we describe the experimental design, data acquisition, and data filtering steps. Next, we present the model using standard statistical notation. Subsequently, we present the results of the inferences based on the model. Finally, we illustrate the usefulness of the model for summarizing complex data and in making predictions given access to different types of preliminary data. We close with discussion and conclusions.

EXPERIMENTAL SECTION

Data. The data were collected by performing 84 different liquid chromatography experiments using a 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 and 35 °C). The detailed conditions are listed in Table S1.

The liquid chromatography experiments were carried out using an Agilent Technologies 1260 Infinity system (Agilent Technologies, Waldbronn, Germany) composed of a binary pump, a membrane degasser, an autosampler, a column thermostat, and a 6224 time-of-flight (TOF) mass spectrometer with a dual electrospray ionization source (Dual ESI) in positive polarity, using an XBridge-C18 (Waters Ltd., Milford, MA, 3 mm × 50 mm, 2.5 μm) column. A drying gas (nitrogen) flow was set to 11 L/min at 350 °C. The nebulizer pressure was set to 50 psi at a capillary voltage of 4000 V. The skimmer, octopole voltage, and fragmentor voltage were set to 65, 750, and 84 V, respectively. The analyses were carried out in scan mode using a 50−1200 m/z range. The reference substances (with masses of 121.050873 and 922.009798) were delivered to the electrospray ionization (ESI) source and monitored to control the mass measurement accuracy. The extra column volume and system dwell volume (V_d) equaled 0.020 and 1.05 mL, respectively. The column hold-up volume (V_0) was 0.266 mL; the flow rate (F) was 0.5 mL/min; and the injection volume was 2 μL. After the end of each gradient, the final content of organic modifier was held for an additional 6 min. A short, 3.5 min, postrun program was performed to maintain the initial conditions of the system and to equilibrate the chromatographic column.

Approximately 5 μmol of each reference substance was weighed and diluted in MeOH to obtain approximately 15 μmol/mL concentrations of each analyte. Undissolved substances were treated with NaOH or formic acid solutions until complete dissolution was achieved. Subsequently, all diluted substances were mixed (using volumes corresponding to 0.3 μmol of each substance) into one sample. The resulting sample was diluted with MeOH 1:499 (v/v) (1 nmol/mL).

Ammonium bicarbonate, ammonium acetate, and ammonium formate were selected as buffers to control the pH of the mobile phase during chromatographic separation. The buffers were prepared at a concentration of 10 mM. 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 prepared buffers were filtered using 0.45 μm nylon filters and degassed.

The pH was measured at 25 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 pH meter electrode was calibrated using a standard aqueous standard. The relationship between pH and the content of organic modifier for various combinations of organic modifier and buffer was experimentally determined prior to the chromatographic analysis. In this setting, pH, and consequently, pK_a values correspond to an absolute pH scale.[18] The obtained data were then described using quadratic equations for each nominal pH, temperature, and organic modifier (36 equations in total)

$$\mathrm{pH}_{m,b,t,j} = \mathrm{pHo}_{m,b,t} + \alpha 1_{m,b,t}\cdot\varphi_j + \alpha 2_{m,b,t}\cdot\varphi_j^2$$

where m denotes the type of organic modifier (1 for MeOH, 2 for ACN), b represents nominal pH (b = 1...9, corresponding to nominal pH of 2.5, 3.3, 4.1, 4.9, 5.8, 6.8, 8.9, 9.7, and 10.5, respectively), t indicates temperature (1−25 and 2−35 °C), j denotes levels of organic modifier content, and $\mathrm{pHo}_{m,b,t}$, $\alpha 1_{m,b,t}$, and $\alpha 2_{m,b,t}$ are regression coefficients specific for a given condition.

The pH measurements and predictions are given in Figure S1. The estimated values of pHo, $\alpha 1$, and $\alpha 2$ for each chromatographic condition were added to the dataset.

Data Extraction Procedure and Data Filtering Step. The MassHunter Profinder B.08.00 (Agilent Technologies, Waldbronn, Germany) was selected to find all of the matches per formula using "Batch Targeted Feature Extraction" (containing 300 predefined masses for each analyte included in the mixture). A match tolerance of +/−20.00 ppm was set for the identification of compounds, and possible ionization adducts of pseudomolecular ions (H^+, NH_4^+) were taken into account. The resulting data were exported as a detailed CSV. All 84 files were then merged and combined with the experimental design data and analyte-specific information.

The data for analysis were restricted to analytes that had "Identification Scores" 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. This process

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

led to the final dataset with 187 analytes (out of the initial 300 analytes).

The molecular structure of the analytes was converted from SMILES format to MDL mol format using OpenBabel.[19] The input molecules were then analyzed for the presence of approximately 204 functional groups and structural elements using Checkmol (version 0.5b N. Haider, University of Vienna, 2003−2018).[20] Functional groups that were not present on any analyte and functional groups merging other simpler functional groups were excluded from the analysis. The lipophilicity (log P), dissociation constant (pK_alit), and predicted error of pK_a (pK_aliterror) were calculated using the ACD/Labs program[21] based on the structures of analytes generated from SMILES strings. Only pK_alits from 2 to 11 were considered. The log P value of the analytes ranged from −4.49 to 7.81, and the molecular mass ranged from 120 to 915. There were 62 and 126 compounds with at least one acidic and basic group. 21 (11.2%) analytes were neutral, 111 (59.4%) analytes were monoprotic and 55 (29.4%) analytes were diprotic in the considered range of pH values. 60 unique functional groups were identified and utilized during the model-building process. The raw data applied in this work are shown in the Supporting Information and graphically presented in Figure S2. The functional groups and their frequency of occurrence are characterized in Figure S3.

Structural Model. A standard chromatographic model was employed in this work.[22,23] The following function describes the relationship between the isocratic retention factor and pH for an analyte with R dissociation steps and $R + 1$ forms[24]

$$ki_{i,\mathrm{m,b,t},j} = \frac{k_{1,i,\mathrm{m,b,t},j} + \sum_{\mathrm{r}=1}^{\mathrm{R}} k_{\mathrm{r}+1,i,\mathrm{m,b,t},j} 10^{\mathrm{r}\cdot \mathrm{pH}_{\mathrm{m,b,t},j} - \sum_{r=1}^{R} pK_{\mathrm{a}_{\mathrm{r},i,\mathrm{m},j}}}}{1 + \sum_{\mathrm{r}=1}^{\mathrm{R}} 10^{\mathrm{r}\cdot \mathrm{pH}_{\mathrm{m,b,t},j} - \sum_{r=1}^{R} pK_{\mathrm{a}_{\mathrm{r},i,\mathrm{m},j}}}} \quad (1)$$

where r represents the dissociation step, i denotes the analyte, m indicates the organic modifier, j represents the organic modifier content, b denotes the pH, and t indicates the temperature. Thus, $pK_{\mathrm{a}\ \mathrm{r},i,\mathrm{m},j}$ denotes the rth dissociation constant of the ith analyte for the mth organic modifier and jth organic modifier content, $k_{\mathrm{r},i,\mathrm{m,b,t},j}$ represents the retention factor of a particular form of the ith analyte in a given chromatographic condition and $ki_{i,\mathrm{m,b,t},j}$ represents the isocratic retention factors in a given chromatographic condition. Furthermore, it was assumed that k depends on the organic modifier content, pH, and temperature according to the following equation

$$\log k_{\mathrm{r,m},i,\mathrm{b,t},j} = \log kw_{\mathrm{r},i} - \frac{S1_{\mathrm{r},i,\mathrm{m}}\cdot(1 + S2_{\mathrm{m}})\cdot\varphi_j}{1 + S2_{\mathrm{m}}\cdot\varphi_j} + d\log kT_i\cdot\frac{T_t - 25}{10} + |\mathrm{chargeA}_{\mathrm{r},i}|\cdot a\mathrm{pHA}\cdot(\mathrm{pH}_{m,\mathrm{b},t,\mathrm{j}} - 7) + |\mathrm{chargeB}_{\mathrm{r},i}|\cdot a\mathrm{pHB}\cdot(\mathrm{pH}_{\mathrm{m,b,t},j} - 7) \quad (2)$$

where $\log kw_{\mathrm{r},i}$ represents the logarithm of retention factors extrapolated to 0% of organic modifier content at 25 °C for mobile phase at pH = 7 for the neutral and dissociated forms of the analyte; $S1_{\mathrm{r},i,\mathrm{m}}$ and $S2_{\mathrm{m}}$ denote slopes in the Neue equation; $d\log kT_i$ denotes the change in log kw due to the increase in temperature by 10 °C, a pH denotes pH effects for cations and anions (common for all analytes); $\mathrm{charge}A_{\mathrm{r},i}$ and $\mathrm{charge}B_{\mathrm{r},i}$ denote a charge state of an analyte ($\mathrm{charge}A_{\mathrm{r},i} = \{0, -1, -2, ...\}$ for anions, and $\mathrm{charge}B_{\mathrm{r},i} = \{0, 1, 2, ...\}$ for cations); and |.| denotes absolute value. In this parametrization of the Neue equation, the $S1$ parameter reflects the difference between the logarithm of retention factors corresponding to water (0% organic modifier content) and MeOH or ACN (100% organic modifier content) as eluents.

Furthermore, a linear relationship between pK_a values and organic modifier content was assumed

$$pK_{\mathrm{a}\ \mathrm{r},i,\mathrm{m},j} = pK_a w_{\mathrm{r},i} + \alpha_{\mathrm{r},i,\mathrm{m}}\cdot\varphi_j \quad (3)$$

where $pK_{\mathrm{a}\ \mathrm{r},i,\mathrm{m},j}$ denotes dissociation constants of an analyte in given chromatographic conditions, $pK_a w_{\mathrm{r},i}$ denotes aqueous pK_a, and $\alpha_{\mathrm{r},i,\mathrm{m}}$ denotes the slope due to changes in the organic modifier. The linear relationship is generally valid for $\varphi_j < 0.8$.

Measurement-Error Model. The observed retention factors ($t_{\mathrm{Robs},z}$) were modeled using the following model

$$t_{\mathrm{Robs},z} \sim \mathrm{student\_}t(\nu, t_{\mathrm{R},z}, \sigma_{i[z]}) \quad (4)$$

where z denotes the zth measurement and student_t denotes the Student's t-distribution with the mean given by the predicted retention time $t_{\mathrm{R},z}$, scale σ_i (analyte-specific), and normality parameter ν. The retention time $t_{\mathrm{R},z}$ under an organic modifier gradient was calculated utilizing the well-known integral equation

$$\int_0^{t_{\mathrm{R},z} - t_0 - t_e} \frac{\mathrm{d}t}{t_0\cdot ki_z(t)} = 1 \quad (5)$$

where t_0 denotes column hold-up (dead) time, t_e denotes extra column time, and $ki_z(t)$ denotes the instantaneous isocratic retention factor corresponding to the mobile phase composition at time t at the column inlet for a particular measurement. 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.

Analyte-Level Model. The $\log kw_{\mathrm{r},i}$ and $S1_{\mathrm{r},i,\mathrm{m}}$ parameters for each analyte form were calculated based on log kw and $S1$ of the neutral form of an analyte and the difference in log kw or $S1$ values between the neutral form of an analyte and the ionized form of an analyte. The $S1$ parameter was separately estimated for MeOH ($m = 1$) and ACN ($m = 2$)

$$\log kw_{\mathrm{r},i} = \log kwN_i + |\mathrm{chargeA}_{\mathrm{r},i}|\cdot d\log kwA_{\mathrm{r},i} + |\mathrm{chargeB}_{\mathrm{r},i}|\cdot d\log kwB_{\mathrm{r},i} \quad (6)$$

$$S1_{\mathrm{r},i,\mathrm{m}=1} = S1mN_i + |\mathrm{chargeA}_{\mathrm{r},i}|\cdot dS1mA_{\mathrm{r},i} + |\mathrm{chargeB}_{\mathrm{r},i}|\cdot dS1mB_{\mathrm{r},i} \quad (7)$$

$$S1_{\mathrm{r},i,\mathrm{m}=2} = S1aN_i + |\mathrm{chargeA}_{\mathrm{r},i}|\cdot dS1aA_{\mathrm{r},i} + |\mathrm{chargeB}_{\mathrm{r},i}|\cdot dS1aB_{\mathrm{r},i} \quad (8)$$

Furthermore, the α parameters were assumed to be different for acids and bases and were separately estimated for MeOH ($m = 1$) and ACN ($m = 2$)

$$\alpha_{\mathrm{r},i,\mathrm{m}=1} = \alpha mA_{\mathrm{r},i}\cdot\mathrm{groupA}_{\mathrm{r},i} + \alpha mB_{\mathrm{r},i}\cdot\mathrm{groupB}_{\mathrm{r},i} \quad (9)$$

$$\alpha_{\mathrm{r},i,\mathrm{m}=2} = \alpha aA_{\mathrm{r},i}\cdot\mathrm{groupA}_{\mathrm{r},i} + \alpha aB_{\mathrm{r},i}\cdot\mathrm{groupB}_{\mathrm{r},i} \quad (10)$$

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

where group$A_{r,i}$ and group$B_{r,i}$ denote the type of dissociating group (group$A_{r,i}$ = 1 if acidic and 0 otherwise; group$B_{r,i}$ = 1 if basic and 0 otherwise).

The second-level part of the model describes the relationship between analyte-specific parameters and predictors. The parameters for the neutral form of an analyte were assumed to be correlated and related to log P and functional groups

$$\begin{bmatrix} \log kwN_i \\ S1mN_i \\ S1aN_i \end{bmatrix} \sim \mathrm{MNV}\left(\begin{bmatrix} \theta_{\log kwN} + \beta_{\log kwN}\cdot(\log P_i - 2.2) + \pi_{\log kwN}\cdot X \\ \theta_{S1mN} + \beta_{S1mN}\cdot(\log P_i - 2.2) + \pi_{S1mN}\cdot X \\ \theta_{S1aN} + \beta_{S1aN}\cdot(\log P_i - 2.2) + \pi_{S1aN}\cdot X \end{bmatrix}, \Omega \right) \tag{11}$$

where MVN denotes the multivariate normal distribution; $\theta_{\log kwN}$, θ_{S1mN}, and θ_{S1aN} are the mean values of individual chromatographic parameters that correspond to a typical analyte with log P = 2.2, with no functional groups at 25 °C; $\beta_{\log kw}$, β_{S1m}, and β_{S1a} are regression coefficients between the individual chromatographic parameters and the log P_i values; and π is an effect of each functional group on chromatographic parameters with separate values for log kwN, $S1mN$, and $S1aN$. 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 187 × 60 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 and Ω denoting a variance-covariance matrix. To ease the specification of the prior distribution, Ω was decomposed into a vector of scales (vector $\omega = [\omega_{\log kwN}, \omega_{S1mN}, \omega_{S1aN}]$) and a correlation matrix (3 × 3 matrix ρ_1) based on the formula

$$\Omega = \mathrm{diag}(\omega)\cdot\rho_1\cdot\mathrm{diag}(\omega) \tag{12}$$

The difference in retention between the ionized form of an analyte and the neutral form of an analyte was separately estimated for acids and bases

$$d\log kwA_{r,i} \sim N(\theta_{d\log kwA}, \kappa_{d\log kw}),\ d\log kwB_{r,i} \sim N(\theta_{d\log kwB}, \kappa_{d\log kw}) \tag{13}$$

$$dS1mA_{r,i} \sim N(\theta_{dS1mA}, \kappa_{dS1m}),\ dS1mB_{r,i} \sim N(\theta_{dS1mB}, \kappa_{dS1m}) \tag{14}$$

$$dS1aA_{r,i} \sim N(\theta_{dS1aA}, \kappa_{dS1a}),\ dS1aB_{r,i} \sim N(\theta_{dS1aB}, \kappa_{dS1a}) \tag{15}$$

where $\theta_{d\log kw}$ and θ_{dS1} denote the mean, and $\kappa_{d\log kw}$ and κ_{dS1} represent the standard deviation for acids (A) and bases (B) in MeOH (a) or ACN (m).

The effect of temperature was assumed to differ for each analyte and to follow a normal distribution

$$d\log kT_i \sim N(\theta_{d\log kT},\ \omega_{d\log kT}) \tag{16}$$

The p$K_a w_{r,i}$ in water was assumed to be equal to the aqueous literature values p$K_a$lit$_{r,i}$ assuming a reported measurement error p$K_a$literror$_{r,i}$

$$\mathrm{p}K_a w_{r,i} \sim N(\mathrm{p}K_a\mathrm{lit}_{r,i}, \mathrm{p}K_a\mathrm{literror}_{r,i}) \tag{17}$$

Furthermore, the α values for MeOH and ACN were assumed to be correlated for acids and bases

$$\begin{bmatrix} \alpha mA_{r,i} \\ \alpha aA_{r,i} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \theta_{\alpha mA} \\ \theta_{\alpha aA} \end{bmatrix}, Tau \right), \begin{bmatrix} \alpha mB_{r,i} \\ \alpha aB_{r,i} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \theta_{\alpha mB} \\ \theta_{\alpha aB} \end{bmatrix}, Tau \right) \tag{18}$$

where $\theta_{\alpha mA}$, $\theta_{\alpha aA}$, $\theta_{\alpha mB}$, and $\theta_{\alpha aB}$ denote the mean α for acids (A) and bases (B) in MeOH (a) or ACN (m). Tau was also decomposed into a vector of scales vector $\tau = [\tau_{\alpha m}, \tau_{\alpha a}]$ and a correlation matrix (2 × 2 matrix ρ_2) based on the following formula

$$Tau = \mathrm{diag}(\tau)\cdot\rho_2\cdot\mathrm{diag}(\tau) \tag{19}$$

Priors. The Bayesian model requires specification of priors that provide a likely range of model parameters expected before the data are observed. Priors also provide the appropriate scales for a given analysis and introduce regularization into the analysis. In this work, the prior information was selected to be weakly informative and in agreement with known facts about analyte retention and gradient chromatography.

The retention factor of the typical neutral form of an analyte ($\theta_{\log kwN}$) was assumed to equal 2.2 ± 2, where 2.2 and 2 correspond to the mean and standard deviation of log P_i values. The typical $S1$ values in MeOH and ACN (θ_{S1mN} and θ_{S1aN}) were assumed to be approximately 4 and 5, respectively, with a standard deviation of 1.[25]

$$\theta_{\log kwN} \sim N(2.2,\ 2),\ \theta_{S1mN} \sim N(4,\ 1),\ \theta_{S1aN} \sim N(5,\ 1) \tag{20}$$

The slope (β) was assumed to be nearly 1 (±0.125) for the log kwN vs log P relationship and to be 0.5 (±0.5) for the $S1$ vs log P relationships

$$\beta_{\log kwN} \sim N(1,\ 0.125),\ \beta_{S1mN},\ \beta_{S1aN} \sim N(0.5,\ 0.5) \tag{21}$$

The regression parameters that describe the effects of substituents were given the priors that assume small effects

$$\pi_{\log kwN,1:60},\ \pi_{S1mN,1:60},\ \pi_{S1aN,1:60} \sim N(0,\ \sigma_\pi) \tag{22}$$

$$\sigma_{\pi\ \log kwN},\ \sigma_{\pi,S1mN},\ \sigma_{\pi,S1aN} \sim N_+(0,\ 0.1) \tag{23}$$

where σ_π is a standard deviation of the individual $\pi_{1:60}$ values for a particular parameter.

The value of $d\log kw$, which was assumed to be nearly −1 (±0.125), corresponds to a typical ratio of the retention factors of the neutral and acidic/basic forms on the order of 10.[26] The difference in $S1$ values between the dissociated forms of acids and bases and the neutral form of an analyte was assumed to be similar on average to that in water

$$\theta_{d\log kwA},\ \theta_{d\log kwB} \sim N(-1,\ 0.125),\ \theta_{dS1mA},\ \theta_{dS1mB},\ \theta_{dS1aA},\ \theta_{dS1aB} \sim N(0,\ 0.5) \tag{24}$$

The priors for parameters that describe the effect of pH on retention for cations and anions were selected to avoid such a relationship

$$a\mathrm{pH} \sim N(0,\ 0.1) \tag{25}$$

The $S2$ parameters in the Neue equation were assumed to be similar for all analytes and analyte forms but were assumed

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

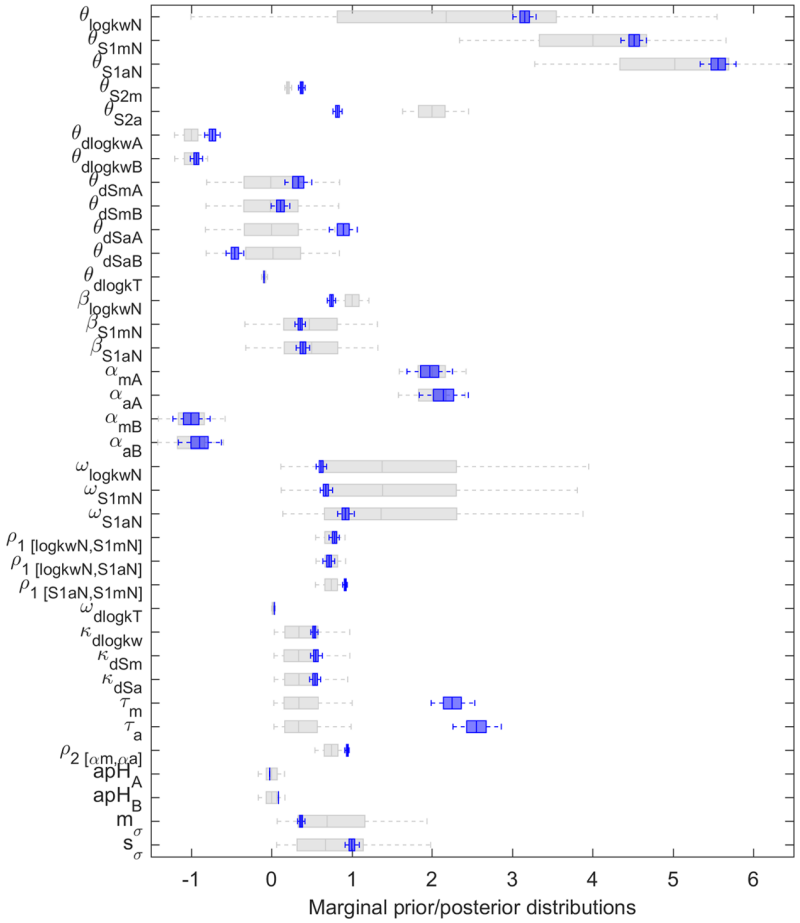


Figure 1. Graphical display of the marginal posterior (blue) and prior (gray) distributions for the population-level parameters. The exact values are given in Table S2.

to differ for each organic modifier. $S2$ was assumed to be positive and nearly 0.2 for MeOH and 2 for ACN.[27]

$$\theta_{S2m} \sim \text{lognormal}\ (\ln(0.2),\ 0.125),\ \ \theta_{S2a} \sim \text{lognormal}\ (\ln(2),\ \ 0.125) \tag{26}$$

According to the literature, acids (such as carboxylic acids and phenol) show modest increases in pK_a (1−2 pK units) in solvent mixtures containing up to 60−70% MeOH and ACN. Basic compounds (such as amines and anilines) display a universal decrease in pK_a (∼1 pK unit) up to approximately 80% organic solvent.[28] Based on these results, it was assumed that pK_a increases with MeOH/ACN content for acids with a typical slope of 2.0 (± 0.125) and that pK_a decreases for bases with a typical slope of −1 (±0.125).

$$\theta_{\alpha mA},\ \theta_{\alpha aA} \sim N(2,\ 0.125),\ \ \theta_{\alpha mB},\ \theta_{\alpha aB} \sim N(-1,\ \ 0.125) \tag{27}$$

Between-analyte variabilities were given weekly informative priors of the following form

$$\omega \sim N_+(0,\ 2),\ \kappa \sim N_+(0,\ 0.5),\ \ \tau \sim N_+(0,\ 0.5) \tag{28}$$

where N_+ denotes the half-normal distribution. The correlation matrix was given the following prior:

$$p(\rho_1) \propto \text{LKJ}(3)\cdot\prod_u N(0.75,\ 0.125) \tag{29}$$

$$p(\rho_2) \propto \text{LKJ}(2)\cdot\prod_u N(0.75,\ 0.125) \tag{30}$$

where ρ_1 and ρ_2 have a joint prior consisting of a uniform LKJ prior (Lewandowski et al. distribution[29]) on the matrix and a normal on its elements.[30] The symbol ∝ means "is proportional to". LKJ(3) and LKJ(2) ensure that the densities are uniform over correlation matrices of order 3 and order 2, respectively, and u denotes the unique lower triangular elements of ρ_1 and ρ_2. High correlations can be expected for $\log kwN_i$, $S1mN_i$, and $S1aN_i$, and similarly, between $\alpha m_{r,i}$ and $\alpha a_{r,i}$. Here, we assumed that all of the correlation coefficients are high and positive.

Priors for the effects of temperature on the retention factor assume a 1−3% decrease in retention factor per unit increase in temperature.[31] This finding corresponds to the following priors

$$\theta_{d\log kT} \sim N(-0.087,\ \ 0.022),\ \ \omega_{d\log kT} \sim N_+(0.022) \tag{31}$$

Priors for residual variability (σ_i, ν) equal

$$\sigma_i \sim \text{lognormal}\ (\ln(m_\sigma),\ \ s_\sigma) \tag{32}$$

$$m_\sigma \sim N_+(0,\ 1),\ \ s_\sigma\ \sim N_+(0,\ 1),\ \ \nu = 3 \tag{33}$$

Under this model, σ_i is lognormally distributed with a typical value of approximately 0.5. Large between-analyte variabilities of σ_i values were assumed. The normality parameter was

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

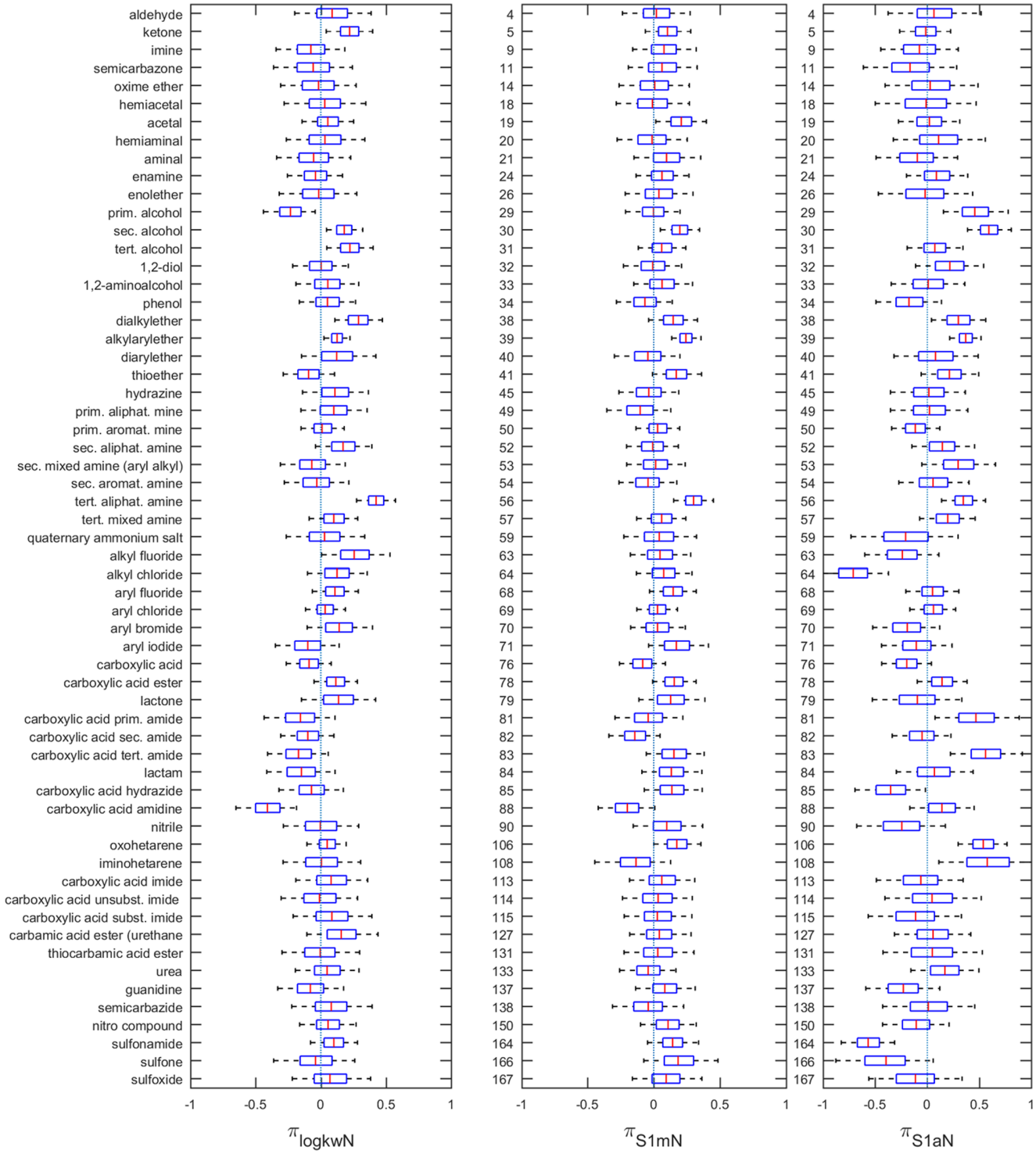


Figure 2. Graphical display of the marginal posterior distributions for the effects of each functional group on log *kwN*, *S1mN*, and *S1aN*.

assumed to equal 3 due to the large number of outlying measurements in the data.

Bayesian Inference. *Technical.* Multilevel modeling was performed in Stan/CmdStan[32] software linked with MATLAB R2017a[33] using MATLAB-Stan 2.15.[34] For the inference and simulation calculations, we used the following values of the Stan parameters: number of iterations = 1000, warmup = 1000, and number of Markov chains = 4. The reduce_sum function, which was selected to accelerate the calculations, works 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 MATLAB code, data, and Stan code used to analyze the data are publicly available from GitHub (https://github.com/wiczling/lcms). The raw data are also available through a repository.[35]

Predictions Using a Limited Set of Experiments. To illustrate the usefulness of the proposed model, we selected six

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

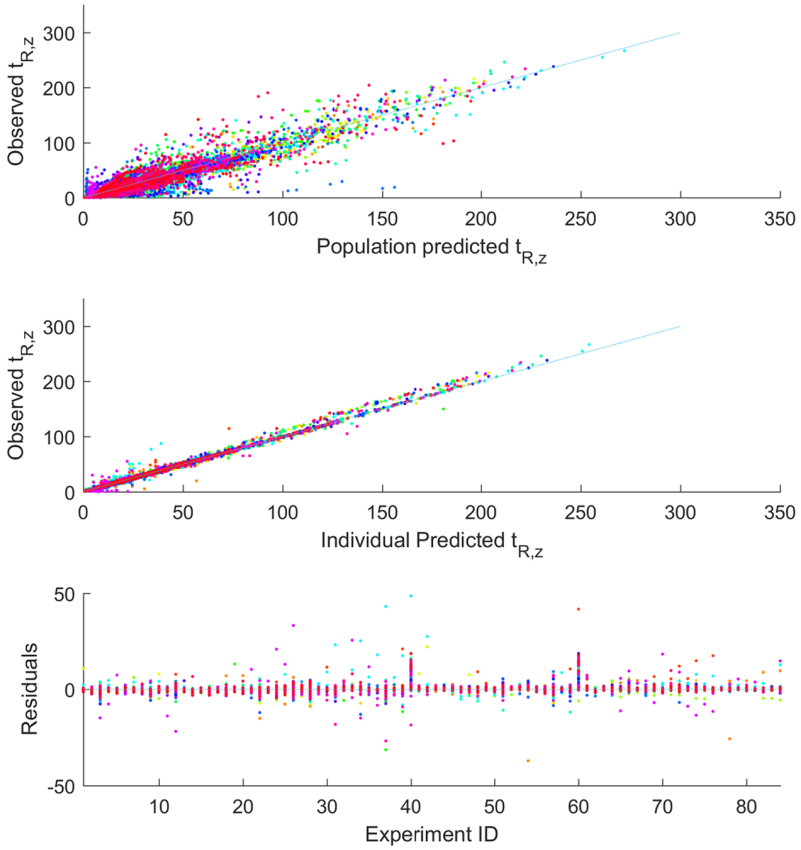


Figure 3. Goodness-of-fit plots. The observed vs the mean population-predicted retention factors (i.e., a posteriori means of predictive distributions corresponding to the future observations of a new analyte), the observed vs the mean individual-predicted retention times (i.e., a posteriori mean of a predictive distribution conditioned on the observed data from the same analyte), and the residuals vs experiment ID.

analytes with different acidic/basic properties (within the range of considered pH values): acridine (monoprotic acid), baclofen (zwitterion: acidic and basic group), nifedipine (neutral), pioglitazone (zwitterion: basic and acidic group), quinine (diprotic: 2 basic groups), and tolbutamide (monoprotic base). The experimental data for these analyses are presented in Figure S4. Three types of predictions are shown: (i) individual predictions that correspond to future observations given access to all of the experimental data collected for these analytes (84 experiments), (ii) population predictions that correspond to future observations when no experimental data are available, and (iii) limited data predictions that correspond to future predictions given access to the limited experimental data collected for these analytes (three experiments collected for pHs of 2.5, 5.8, 10.5, and 30 min MeOH gradient at 25 °C). The predictions are 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.

RESULTS AND DISCUSSION

The available data were described using a single mechanistic model. This model was built using simple blocks/components, known fundamentals of gradient chromatography, and prior knowledge available in the literature. This approach allowed us to build a fairly realistic model using simple and interpretable parameters (such as $\log kw$, S, and pK_a). However, we are aware that the results of this work are sensitive to the choice of priors. For this reason, we explicitly described our choice and encouraged readers to criticize it and modify it according to a different state of knowledge about the problem.

Figure 1 (Table S1) shows a summary of the marginal posterior distributions for population-level parameters compared with an assumed prior knowledge. The effects of functional groups are shown in Figure 2. These population-level parameters summarize the behavior of a typical analyte and contain the information required to predict retention for a new analyte. The developed model generated chromatographic parameters that generally show agreement with the literature knowledge and assumed priors. It is expected, as many heuristics about chromatographic parameters are available in the literature (e.g., $S1$ is nearly 4 for methanol, or unit increase in the temperature decrease retention factor by 1−3%). The most surprising finding (difference between prior/posterior distribution) was observed for the $dS1a$, $S2$, and τ parameters. The $dS1a$ parameters describe the retention of anions/cations in the ACN-rich mobile phase. It was observed that $\theta_{dS1aA} = 0.89\ (0.72 - 1.07)$ and that $\theta_{dS1aB} = -0.46((-0.57) - (-0.35))$. These parameters have different signs, suggesting

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

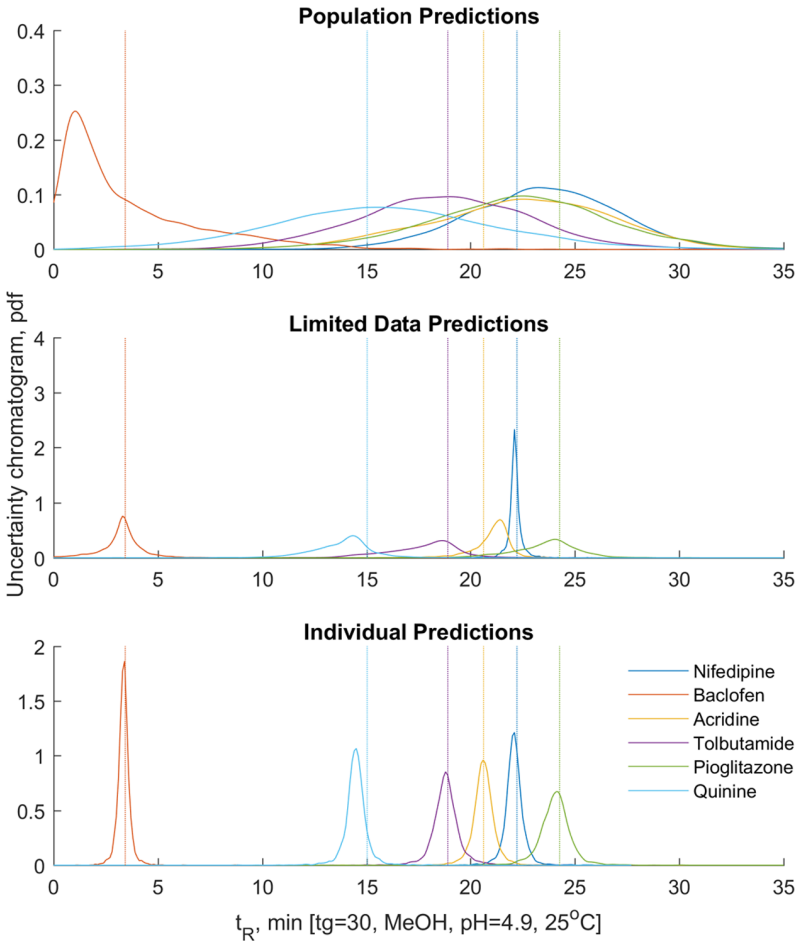


Figure 4. Uncertainty chromatograms displaying the predictions for six selected analytes using different preliminary information. Each peak represents the range of analyte retention factors compatible with prior and preliminary data. Predictions were based on three experiments conducted at pH values of 2.5, 5.8, and 10.5 for a 30 min MeOH gradient at 25 °C. Colors correspond to different analytes that are identified in the bottom subplot. Vertical lines represent actual measurements.

different retention characteristics of anions and cations in ACN-rich mobile phases. This difference is less evident for MeOH, as θ_{dS1mA} = 0.33 (0.17 − 0.50) and θ_{dS1mB} = 0.11 ((−0.01) − 0.23). Between-analyte variability for α is large (τ_m = 2.26 (1.99 − 2.53) and τ_a = 2.56 (2.26 − 2.86)). To some degree, it might be a consequence of misidentification of certain analytes and consequential error in the value of predictors (pK_a, log P, charges, and groups). The $S2$ parameters are higher for MeOH and lower in ACN than expected. It was also evident that the stationary phase changes its properties with the pH of the mobile phase (or buffer type). In this work, this change was quantitated by determining the slope of the relationship between log k_w and pH. The slope is negative for anions −0.02((−0.03) − (−0.02)) and positive for cations (0.09 (0.08 − 0.09)). This effect is likely caused by a combination of various mechanisms related to the pH of the mobile phase, such as the presence of surface silanol groups and the formation of ion pairs with buffer components.[26,36]

Figures S5−S7 visualize the distribution of analyte-specific (individual) chromatographic parameters and their relationship to predictors (log P and pK_a). There are strong correlations between individual parameters corresponding to neutral forms of analytes: 0.78 (0.71 − 84), 0.71 (0.64 − 0.78), and 0.92 (0.88 − 0.94) for log kwN−$S1mN$, log kwN−$S1aN$, and SmN−$S1aN$ correlations, respectively. Additionally, the α values for MeOH and ACN are highly correlated (the correlation is 0.94 (0.91 − 0.96)). A high correlation implies mutual information between the variables. Simply, one can gain knowledge about one parameter by knowing the value of another parameter (e.g., the knowledge about log kwN_i narrows the range of possible $S1mN_i$ values for a particular analyte). This finding is confirmed in practice, as one scouting gradient run (assuming MeOH or ACN content as a single design variable) usually provides much information about retention. This result is a consequence of a high correlation between log kw and $S1$, which implies that one "effective" parameter drives retention.

It would be valuable to identify chromatographic parameters that are approximately independent of the stationary phase. Such parameters (if precisely and accurately identified) can serve as a prior for data predictions for other stationary phases (chromatographic columns). The parameters meeting this criterion are pK_a and α (they reflect acid−base properties of compounds in the solvent). Additionally, $S1$ and $dS1$ seem to describe the properties of the mobile phase. $S1$ represents the difference in retention between log kw and log km (or log ka). This parameter reflects the total free energy of transfer of the compound from water to MeOH or ACN. $dS1$ corresponds to a change in $S1$ due to dissociation. In contrast, the parameters that strongly depend on stationary phase characteristics are

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

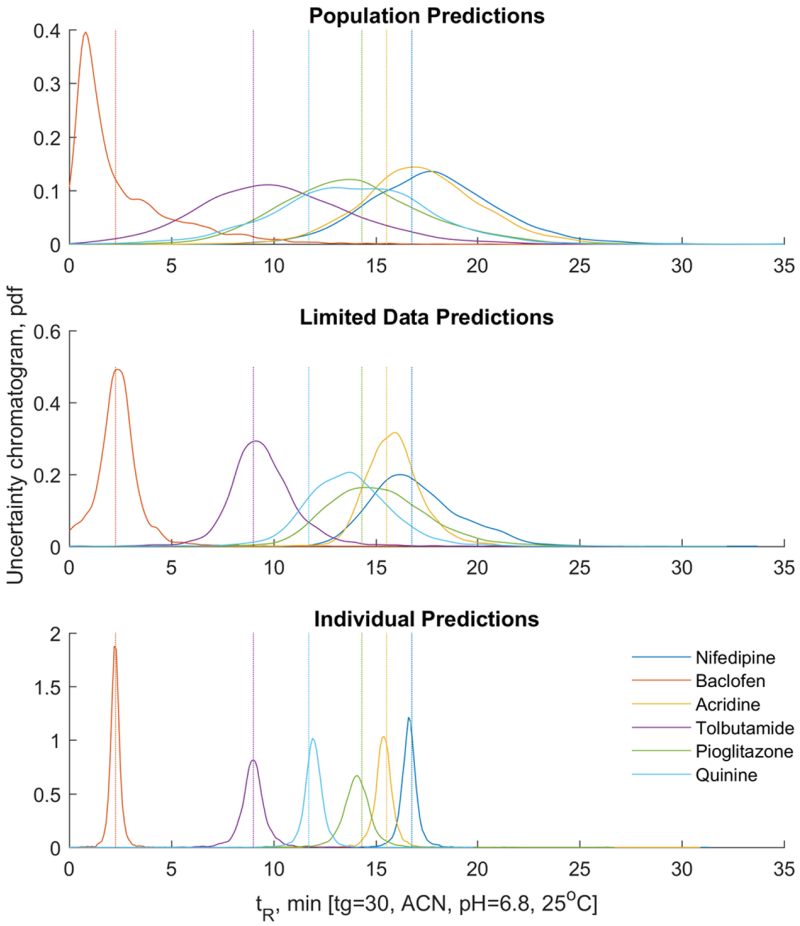


Figure 5. Uncertainty chromatograms displaying the predictions for six selected analytes using different preliminary information. Each peak represents the range of analyte retention factors compatible with prior and preliminary data. Predictions were based on three experiments conducted at pH values of 2.5, 5.8, and 10.5 for a 30 min MeOH gradient at 25 °C. Colors correspond to different analytes that are identified in the bottom subplot. Vertical lines represent actual measurements.

log *kwN*, *d*log *kwN*, $\pi_{\log kwN}$, and *a*pH. The question remains whether they can be accurately estimated based on the experimental design used in this work.

Various interactions occurring in a chromatographic column are usually characterized by a carefully designed set of experiments using probe analytes. This process allows for a detailed physical description of the chromatographic systems.[37] One can also characterize these interactions using retention time data collected for a relatively large and heterogeneous group of compounds chromatographed in a broad range of conditions. Nevertheless, population and individual parameters obtained this way should be treated as "macro" parameters (parameters that describe general behaviors of analytes in the column). As pointed out by Gritti and Guiochon,[37] such "macro" parameters (e.g., log *k*) mask the physical reality of various interactions in the chromatographic column, as they lump the consequences of different effects in one parameter. We agree with this statement. Nevertheless, we suggest that these parameters can be interpreted (under the assumed model) and used for predictions.

The model predictions are well calibrated with the data, as shown in Figure 3. The considerable number of outlying measurements was handled using robust residual error. Figure 3 also summarizes the calibration and sharpness of predictions[38] 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. 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). In scenarios with limited preliminary data, one can expect the calibration and sharpness to fall between those observed for population predictions and those observed for individual predictions. The prediction for 6 selected analytes is illustrated in Figures S8–S10. As expected, the individual predictions are highly accurate because they are based on population-level parameters, predictors, and all of the observed retention time measurements. In contrast, the population predictions are rather imprecise, as predictions are based on population-level parameters and predictors. Access to a set of preliminary experiments reduces uncertainty in predictions. This decrease depends on the choice of experiments and analyte characteristics. The comparison of population, individual, and limited data predictions for two chromatographic conditions are shown in Figures 4 and 5. Three preliminary experiments conducted in MeOH allowed prediction of retention in MeOH and ACN and for other pH values. In all cases, the propagated uncertainty can be applied for decision making, e.g., to determine whether achieving the desired separation is feasible.

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

In our opinion, the quantification of uncertainty is virtually missing in method development. However, it is a crucial element for the more realistic use of chromatographic models in practice, especially in problems involving limited preliminary data. In this case, all of the decisions regarding further analytical steps are made under uncertainty.

The multilevel models can also be beneficial in supporting MS identification of unknown compounds by LC/MS analysis. The probability that a particular peak corresponds to a given analyte can be refined by adding information from the observed retention time. The Bayesian approach seems suitable for these types of problems, as for this problem, predictions need to be made under uncertainty. Since the prediction accuracy of current models (that are based on analyte structure) is rather limited, one can also expect that a limited amount of predictive information from retention time measurment will be added to the probability of correct identification/annotation. The methods using Bayesian algorithms for peak detection and compound screening in LC/MS databases are already available in the literature and can be combined with the proposed model.[39,40]

In this work, we proposed a mechanistic model to describe the retention data of 187 small molecules obtained for a wide range of chromatographic conditions. We are fully aware that the proposed model can be improved by adding several omitted complexities (e.g., effects of temperature or the effects of ionic strength on certain parameters). We also acknowledge the complexity and very time-consuming nature of the approach. Nevertheless, even in the current form, the model provides a step forward in the process of finding a general mechanistic model that is applicable to describe chromatographic retention of a heterogeneous group of analytes. We plan to apply the model to different columns to better understand the column variations in the model parameters. We also suggest that the model has value in improving the accuracy and precision of parameter estimation (due to prior information in subsequent analysis) and in providing the necessary input to identify better experimental designs.

CONCLUSIONS

This work provides a pilot study that aims to demonstrate the application of a Bayesian multilevel model to describe the retention time data collected for 187 analytes for a wide range of chromatographic conditions. The analysis characterizes the chromatographic retention of neutral, acidic, and basic analytes. The model is interpretable and provides a compact summary of complex data. The model can be used to predict retention uncertainty based on various numbers of preliminary experiments, and as such, can be useful for decision making under uncertainty. The model also provides prior information for subsequent analysis.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c02034.

Experimental design (Table S1); summary of the MCMC simulations of the marginal posterior distributions of population-level model parameters (Table S2); pH measurements (Figure S1); raw data (Figure S2); functional groups identified by Checkmol (Figure S3); raw data for a set of six analytes (Figure S4); individual parameters (Figures S5−S7); population predictions (Figure S8); limited data predictions (Figure S9); and individual predictions (Figure S10) (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*

Julia Jacyna − *Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, 80-416 Gdańsk, Poland*

Wiktoria Struck-Lewicka − *Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, 80-416 Gdańsk, Poland*

Michał J. Markuszewski − *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.2c02034

Author Contributions

Ł.K., J.J., and W.S.L. collected the experimental data; Ł.K. and A.K. prepared the data for analysis; A.K. and P.W. analyzed the data; P.W. and A.K. wrote the paper with input from all authors; P.W. conceived of the presented idea, designed the study, and supervised the project; and M.M. helped supervise the project.

Funding

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

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Gritti, F. *Anal. Chem.* **2021**, *93*, 5653−5664.

(2) Bouwmeester, R.; Martens, L.; Degroeve, S. *Anal. Chem.* **2019**, *91*, 3694−3703.

(3) Muteki, K.; Morgado, J. E.; Reid, G. L.; Wang, J.; Xue, G.; Riley, F. W.; Harwood, J. W.; Fortin, D. T.; Miller, I. J. *Ind. Eng. Chem. Res.* **2013**, *52*, 12269−12284.

(4) Talebi, M.; Schuster, G.; Shellie, R. A.; Szucs, R.; Haddad, P. R. *J. Chromatogr. A* **2015**, *1424*, 69−76.

(5) Golmohammadi, H.; Dashtbozorgi, Z.; Heyden, Y. V. *Chromatographia* **2014**, *78*, 7−19.

(6) Daghir-Wojtkowiak, E.; Wiczling, P.; Bocian, S.; Kubik, Ł.; Kośliński, P.; Buszewski, B.; Kaliszan, R.; Markuszewski, M. J. *J. Chromatogr. A* **2015**, *1403*, 54−62.

(7) Hancock, T.; Put, R.; Coomans, D.; Heyden, Y. V.; Everingham, Y. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 185−196.

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX

(8) Haddad, P. R.; Taraji, M.; Szücs, R. *Anal. Chem.* **2021**, *93*, 228−256.
(9) Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. *Metabolites* **2020**, *10*, No. 243.
(10) 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.
(11) Yamamoto, Y.; Yajima, T.; Kawajiri, Y. *Chem. Eng. Res. Des.* **2021**, *175*, 223−237.
(12) Wiczling, P.; Kamedulska, A.; Kubik, Ł. *Anal. Chem.* **2021**, *93*, 6961−6971.
(13) Wiczling, P.; Kubik, Ł.; Kaliszan, R. *Anal. Chem.* **2015**, *87*, 7241−7249.
(14) Wiczling, P. *Anal. Bioanal. Chem.* **2018**, *410*, 3905−3915.
(15) He, Q.-L.; Zhao, L. *Sep. Purif. Technol.* **2020**, *246*, No. 116856.
(16) Wiczling, P. *Sep. Sci. plus* **2018**, *1*, 63−75.
(17) Gelman, A.; Vehtari, A.; Hill, J., *Regression and Other Stories*; Cambrige Universiy Press: Cambridge, 2020.
(18) Rosés, M. *J. Chromatogr. A* **2004**, *1037*, 283−298.
(19) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, No. 33.
(20) Haider, N. *Molecules* **2010**, *15*, 5079−5092.
(21) ACD/Labs, *Release 12.0*; Advanced Chemistry Development Inc.: Toronto, ON, Canada, 2011.
(22) Nikitas, P.; Pappa-Louisi, A. *J. Chromatogr. A* **2009**, *1216*, 1737−1755.
(23) Nikitas, P.; Pappa-Louisi, A. *J. Chromatogr. A* **2002**, *971*, 47−60.
(24) Jano, I.; Hardcastle, J. E.; Zhao, K.; Vermillion-Salsbury, R. *J. Chromatogr. A* **1997**, *762*, 63−72.
(25) Téllez, A.; Rosés, M.; Bosch, E. *Anal. Chem.* **2009**, *81*, 9135−9145.
(26) Neue, U. D.; Phoebe, C. H.; Tran, K.; Cheng, Y. F.; Lu, Z. *J. Chromatogr. A* **2001**, *925*, 49−67.
(27) Pappa-Louisi, A.; Nikitas, P.; Balkatzopoulou, P.; Malliakas, C. *J. Chromatogr. A* **2004**, *1033*, 29−41.
(28) Cox, B. G. *Org. Process Res. Dev.* **2015**, *19*, 1800−1808.
(29) Lewandowski, D.; Kurowicka, D.; Joe, H. *J. Multivar. Anal.* **2009**, *100*, 1989−2001.
(30) Martin, S. R. *Informative Priors for Correlation Matrices: An Easy Approach*, 2021, http://srmart.in/informative-priors-for-correlation-matrices-an-easy-approach/.
(31) Snyder, L. R.; Kirkland, J. J.; Glajch, J. L., *Practical HPLC Method Development*; Wiley: New York, 1997.
(32) Stan Development Team, Stan Modeling Language Users Guide and Reference Manual, VERSION. https://mc-stan.org//mc-stan.org/users/citations/ (accessed Jan 18 , 2022).
(33) MATLAB, *9.2.0.556344 (R2017a)*; The MathWorks Inc.: Natick, MA, 2018.
(34) MatlabStan, https://github.com/brian-lau/MatlabStan (accessed Jan 24 , 2022).
(35) Kubik, Ł.; Jacyna, J.; Struck-Lewicka, W.; Markuszewski, M.; Wiczling, P. LC-TOF-MS Data Collected for 300 Small Molecules *XBridge-C18 Column*. 2022.
(36) Méndez, A.; Bosch, E.; Rosés, M.; Neue, U. D. *J. Chromatogr. A* **2003**, *986*, 33−44.
(37) Gritti, F.; Guiochon, G. *Anal. Chem.* **2005**, *77*, 4257−4272.
(38) Gneiting, T.; Balabdaoui, F.; Raftery, A. E. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **2007**, *69*, 243−268.
(39) Woldegebriel, M.; Gonsalves, J.; van Asten, A.; Vivó-Truyols, G. *Anal. Chem.* **2016**, *88*, 2421−2430.
(40) Woldegebriel, M.; Vivó-Truyols, G. *Anal. Chem.* **2015**, *87*, 7345−7355.

https://doi.org/10.1021/acs.analchem.2c02034
Anal. Chem. XXXX, XXX, XXX−XXX