Supporting Information for:

Analysis of Isocratic Chromatographic Retention Data using Bayesian Multilevel Modeling

*Łukasz Kubik, Roman Kaliszan, Paweł Wiczling**

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

*Corresponding author's e-mail: wiczling@gumed.edu.pl

Table of content:

1. List of 58 analyzed substances:

1,3,5-triisopropylbenzene, 1,4-dinitrobenzene, 1,4-naphthaquinone, 1-bromonaphthalene, 1-chloroanthraquinone, 1-methyl-2-pyrrolidinone, 1-naphthylacetic acid, 1-naphthylacetonitrile, 2,2c-dipyridil, 2,2-dinaphthyl ether, 2,4,6-trichloroaniline, 2-chloroaniline, 2-chloropyridine, 2-naphthol, 3,5-dichlorophenol, 3-cyanopyridine, 3-nitrophthalic anhydride, 3-trifluoromethylphenol, 4,4c-bipyridine, 4-aminophenol, 4-chlorophenol, 4-cyanophenol, 4-iodophenol, 4-nitrobenzoic acid, 4-nitrophenol, 4-phenylphenol, acridone, aniline, anisole, anthracene, anthraquinone, benzamide, benzene, benzoic acid, benzonitrile, benzyl chloride, biphenyl, caffeine, carbazole, chlorobenzene, cyclohexanone, dibenzothiophene, hexachlorobutadiene, hydroquinone, indazole, indole, isopropylbenzene, naphthalene, n-hexylbenzene, nicotinamide, nicotinic acid, nitrobenzene, phenanthrene, phenol, phenylhydrazine, pyrene, toluene, xanthene.

2. Calculation of WAIC, LDPCV and RMSECV

The Watanabe-Akaike information criterion (WAIC) and cross-validation summarized as log pointwise predictive density and RMSE were used to estimate out-of-sample prediction accuracy from a fitted Bayesian model.

Root mean square error (RMSE) was calculated using the standard equation: sqrt(E(E(logk) – $logk_{Obs}$)^2), where E() denotes expectation. The cross validated RMSE (RMSECV) was calculated based on predicted logk values from the combined cross-validation subsamples. RMSECV is less appropriate for models that are far from the normal distribution, that is why we decided to use log pointwise predictive density (LPD) of cross validation.

LPD is the sum across observations of the log of the mean of the likelihoods of each chromatographic measurement from the posterior distribution of the parameters. Mathematically it is given by:

$$LPD = \sum_{z=1}^{Nobs} log\left(\frac{1}{S}\sum_{s=1}^{S} p(logk_{Obs,z}|\theta^s)\right)$$

where s =1…S denotes posteriors samples, and θ denotes all model parameters. LPDCV was calculated using the likelihoods of each measurements obtained from the combined cross-validation subsamples.

WAIC is related to LPD: WAIC = -2 (LPD-P), where P is an estimate of the effective number of parameters in the model. It approximates the leave-one-measurement-out cross validation. Please note that WAIC is conceptually similar to Akaike information criterion (AIC) and Bayesian information criterion (BIC). For more detail please see: Vehtari, A., Gelman, A. & Gabry, J. Stat Comput (2017) 27: 1413. https://doi.org/10.1007/s11222-016-9696-4

2. Tables and Figures referenced in the manuscript

Table S-1. Summary of the MCMC simulations of the marginal posterior distributions of the model parameters. For the Unpooled model the second-level parameters were not estimated and are not provided in the table.

Model:	Pooled			Pooled-log P			Pooled-Mmol		
Parameter	Mean	5%	95%	Mean	5%	95%	Mean	5%	95%
θ_{logkw}	3.40	3.00	3.80	3.20	3.00	3.40	3.40	3.20	3.70
θ_{logkm}	-0.64	-0.73	-0.55	-0.71	-0.76	-0.65	-0.66	-0.74	-0.58
θ_{logka}	-0.48	-0.56	-0.41	-1.00	-1.30	-0.76	-0.49	-0.55	-0.43
θ_{lnS2M}	-1.10	-1.40	-0.84	-0.52	-0.56	-0.48	-1.10	-1.30	-0.81
θ_{lnS2A}	0.34	0.19	0.49	0.39	0.24	0.54	0.37	0.22	0.52
σ	0.033	0.032	0.034	0.033	0.032	0.034	0.03	0.032	0.034
$\rho_{logkm,logka}$	0.84	0.74	0.91	0.53	0.29	0.71	0.76	0.62	0.86
$\rho_{logkm,\ lnS2M}$	0.00	-0.25	0.25	0.17	-0.08	0.41	-0.22	-0.47	0.05
$\rho_{logkm,\ lnS2A}$	0.05	-0.18	0.29	-0.09	-0.32	0.14	-0.30	-0.53	-0.04
$\rho_{logkm,logkw}$	0.78	0.67	0.86	0.24	-0.01	0.46	0.71	0.56	0.83
$\rho_{logka,\ lnS2M}$	-0.18	-0.43	0.07	-0.11	-0.36	0.14	-0.42	-0.63	-0.17
$\rho_{logka,\ lnS2A}$	0.11	-0.14	0.34	0.07	-0.18	0.31	-0.18	-0.44	0.09
$\rho_{logka,\ logkw}$	0.79	0.68	0.87	0.24	0.00	0.46	0.74	0.60	0.85
$\rho_{lnS2M,\ lnS2A}$	0.76	0.62	0.87	0.75	0.60	0.86	0.74	0.58	0.86
$\rho_{lnS2M,\ logkw}$	0.21	-0.06	0.45	0.73	0.58	0.85	-0.15	-0.42	0.13
$\rho_{lnS2A,\ logkw}$	0.41	0.20	0.60	0.73	0.60	0.84	-0.01	-0.28	0.26
ω_{logkw}	0.35	0.29	0.43	0.21	0.16	0.26	0.31	0.25	0.37
ω_{logkm}	0.28	0.23	0.34	0.15	0.11	0.18	0.24	0.19	0.28
ω_{logka}	0.90	0.69	1.10	0.84	0.65	1.10	0.92	0.71	1.20
ω_{lnS2M}	0.56	0.44	0.70	0.54	0.42	0.66	0.57	0.45	0.71
ω_{lnS2A}	1.50	1.20	1.80	0.59	0.46	0.74	0.88	0.69	1.10
$\beta_{1,logkw}$				1.30	1.20	1.40			
$\beta_{1,logka}$				0.21	0.18	0.25			
$\beta_{1,logkm}$				0.29	0.25	0.33			
$\beta_{2,logkw}$							0.029	0.024	0.035
$\beta_{2,logka}$							0.0043	0.0030	0.0056
$\beta_{2,logkm}$							0.0058	0.0040	0.0076
ν	2.90	1.90	4.20	2.60	1.70	3.80	2.90	1.90	4.20

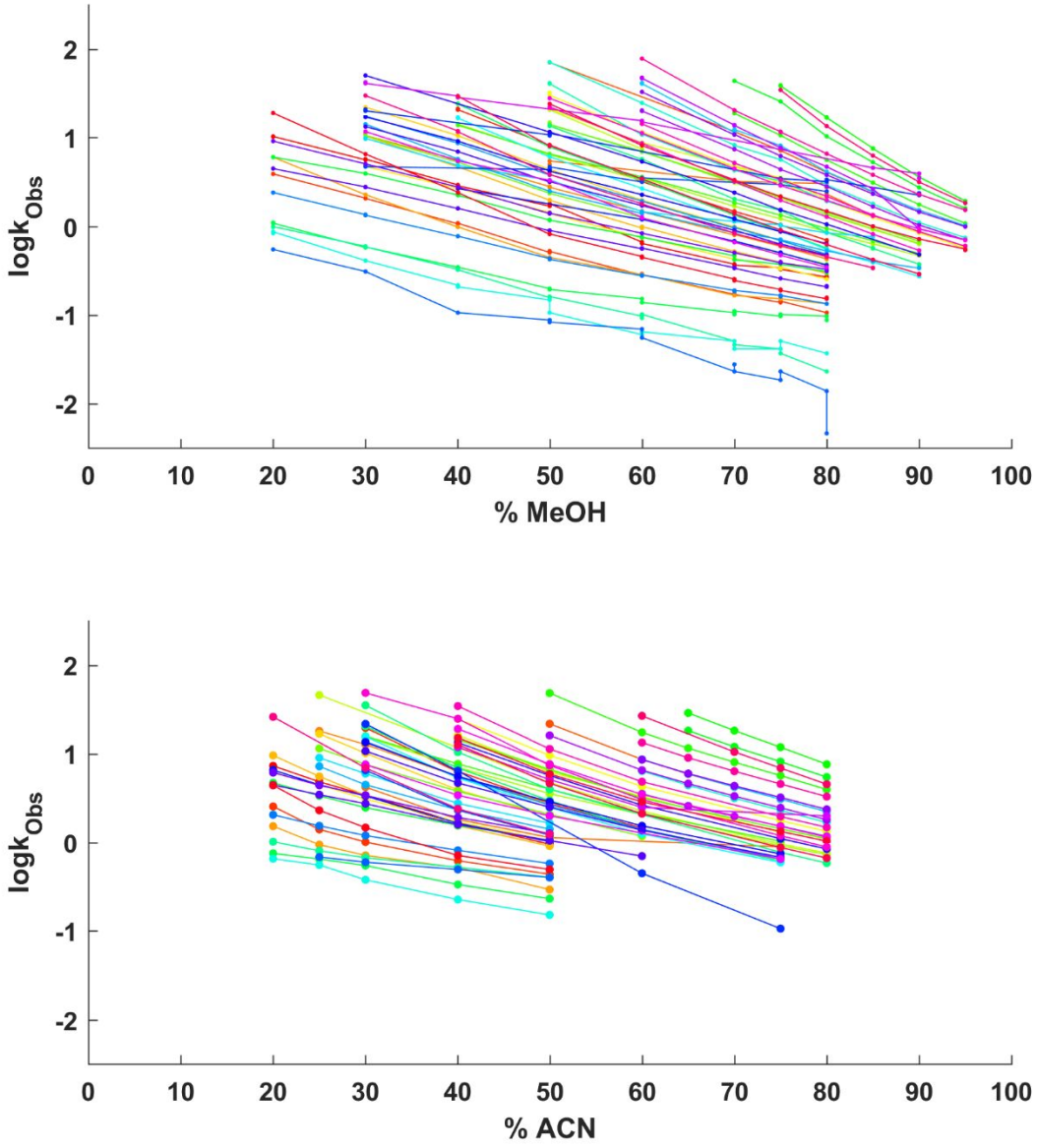
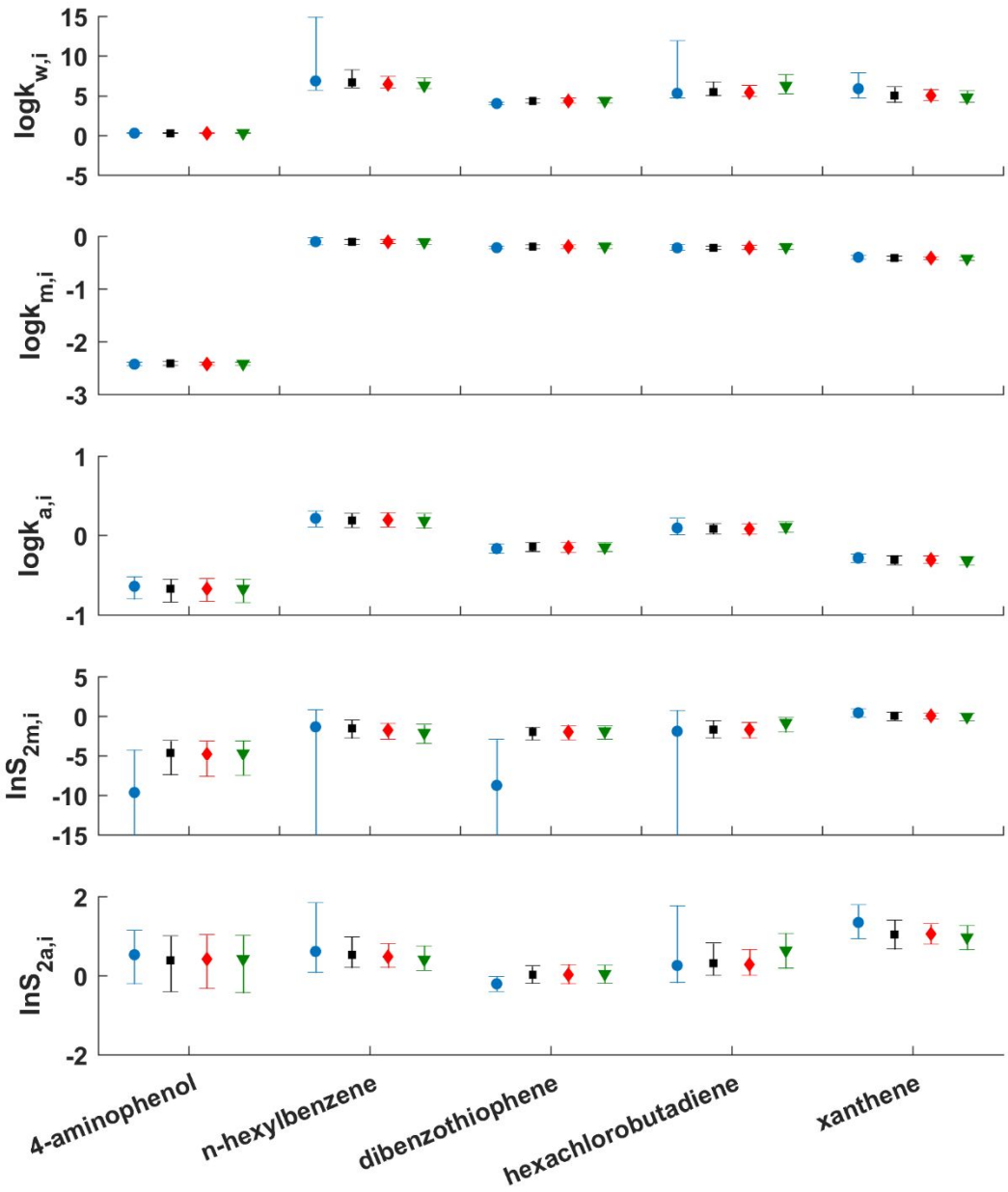


Figure S-1. The graphical representation of the raw data. The retention factors were obtained under the isocratic conditions using MeOH/Water (upper plot) and ACN/Water (lower plot) mobile phases.



60

61 **Figure S-2.** Chromatographic parameters (posterior median and 90% (5^{th}-95^{th} percentile) credible
62 intervals) for 5 representative analytes obtained based on the Unpooled (blue), Pooled (black),
63 Pooled-log P (red), Pooled-Mmol (green) models. The shrinkage is clearly evident for $lnS_{2a,i}$,
64 $lnS_{2m,i}$ and log $k_{w,i}$ parameters.

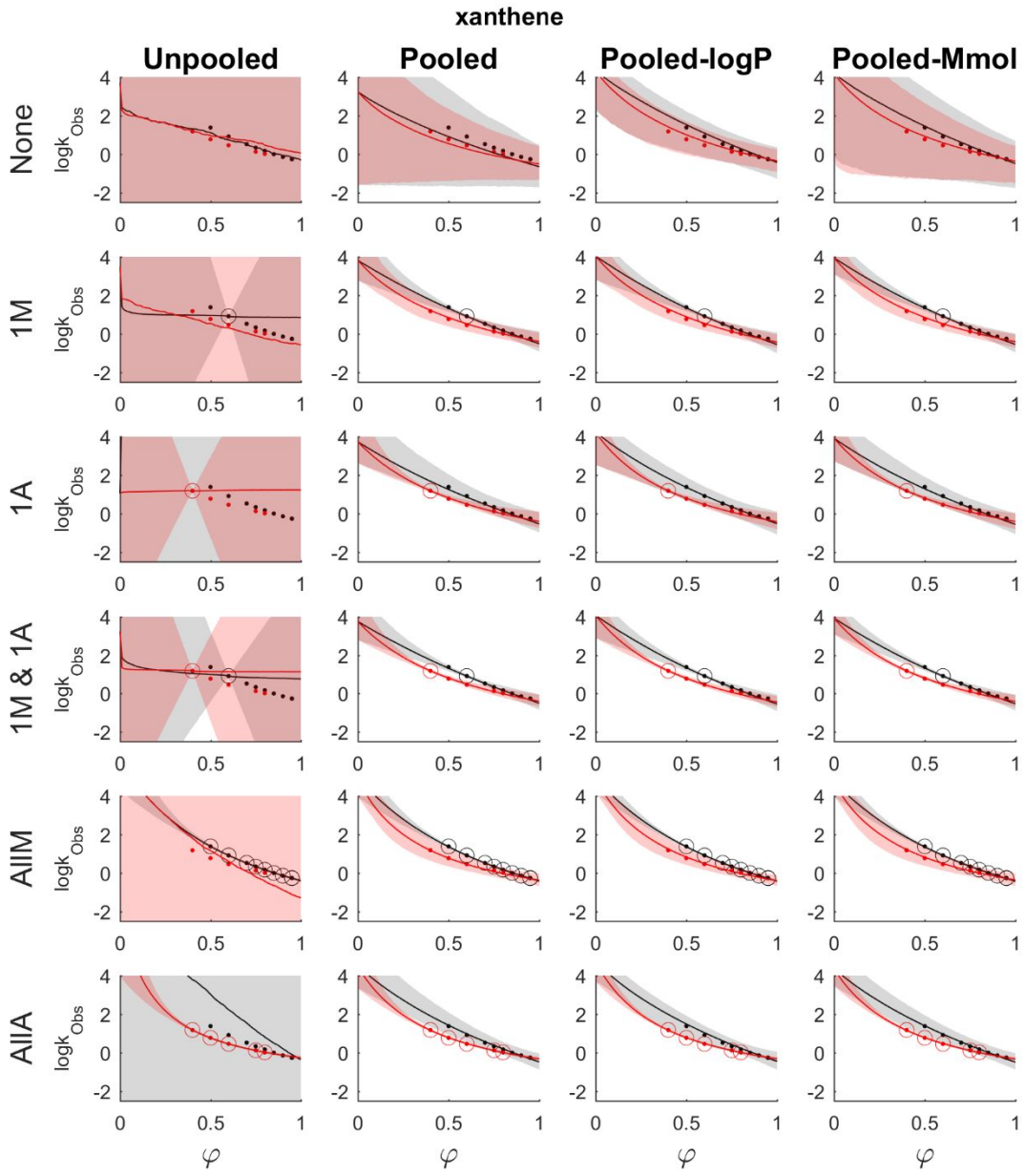


Figure S-3. Predicted (posterior median (line) and 95% credible intervals (shaded area)) and observed retention factors (dots) for xanthene. Black color corresponds to MeOH whereas red color corresponds to ACN. Prediction corresponds to the future observations on the same analyte condition on the part of the observed data (specifically none, single MeOH (1M), single ACN (1A), single MeOH and single ACN (1M & 1A), All MeOH (AllM) and all ACN (AllA)) from the same analyte. The data used for parameter estimation is given as open circles.

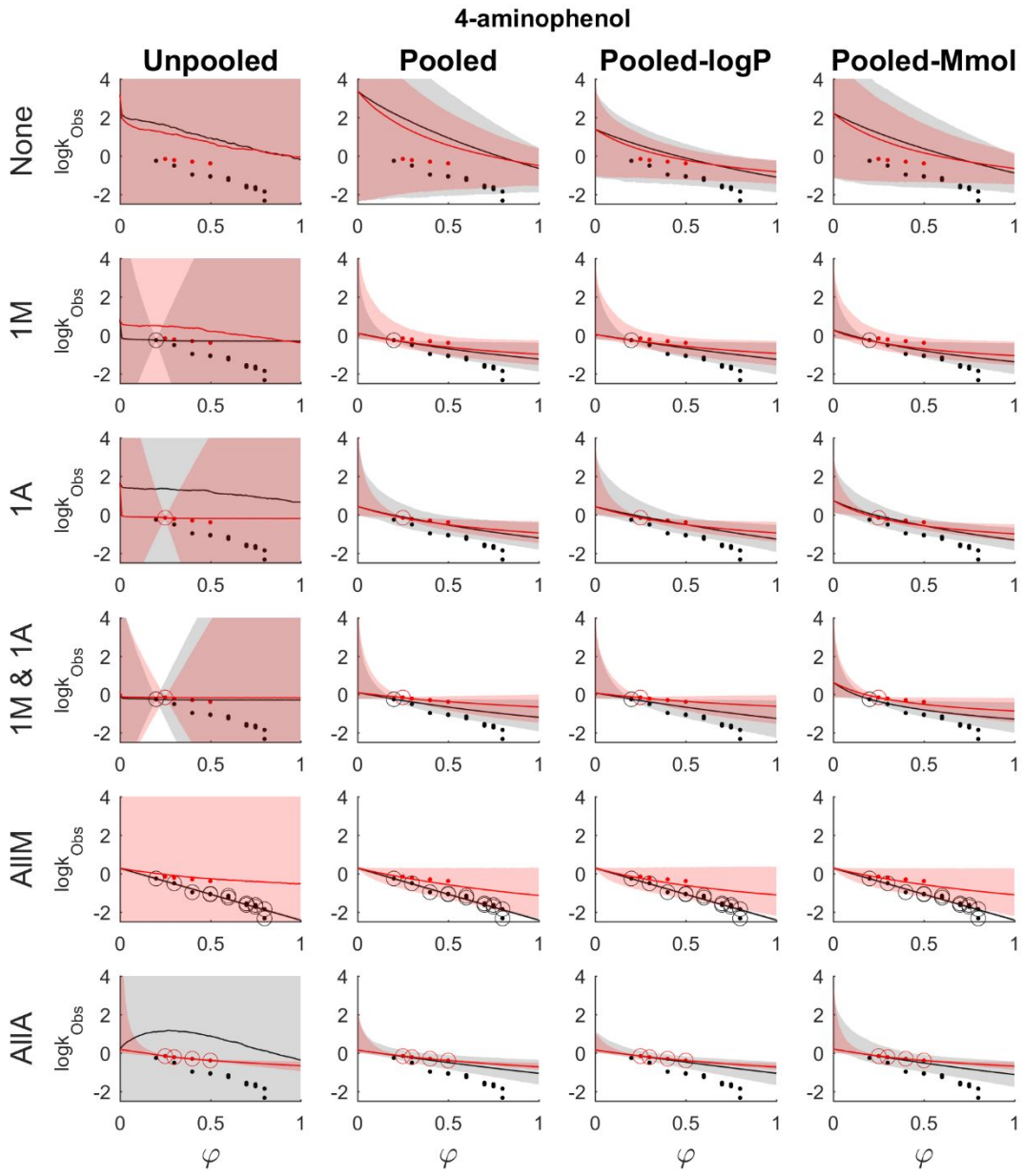


Figure S-4. Predicted (posterior median (line) and 95% credible intervals (shaded area)) and observed retention factors (dots) for 4-aminophenol. Black color corresponds to MeOH whereas red color corresponds to ACN. Prediction corresponds to the future observations on the same analyte condition on the part of the observed data from the same analyte (specifically none, single MeOH (1M), single ACN (1A), single MeOH and single ACN (1M & 1A), All MeOH (AllM) and all ACN (AllA)). The data used for parameter estimation is given as open circles.

3. The excerpt from the attached data file (Data.csv)

Name	ID	MeOH1ACN2	fi	logkObs	MLOGP	MMOL
4-cyanophenol	1	1	0.8	-0.58058	1.155988	119.13
4-cyanophenol	1	1	0.8	-0.56561	1.155988	119.13
4-cyanophenol	1	1	0.8	-0.57303	1.155988	119.13
4-cyanophenol	1	1	0.75	-0.4852	1.155988	119.13
4-cyanophenol	1	1	0.75	-0.4614	1.155988	119.13
4-cyanophenol	1	1	0.75	-0.46723	1.155988	119.13
4-cyanophenol	1	1	0.7	-0.43883	1.155988	119.13
4-cyanophenol	1	1	0.7	-0.43883	1.155988	119.13
4-cyanophenol	1	1	0.7	-0.42797	1.155988	119.13
4-cyanophenol	1	1	0.6	-0.19344	1.155988	119.13
4-cyanophenol	1	1	0.6	-0.18112	1.155988	119.13
4-cyanophenol	1	1	0.6	-0.1781	1.155988	119.13
4-cyanophenol	1	1	0.5	0.254605	1.155988	119.13
4-cyanophenol	1	1	0.5	0.230567	1.155988	119.13
4-cyanophenol	1	1	0.5	0.228206	1.155988	119.13
4-cyanophenol	1	1	0.4	0.456632	1.155988	119.13
4-cyanophenol	1	1	0.4	0.466997	1.155988	119.13
4-cyanophenol	1	1	0.3	0.757312	1.155988	119.13
4-cyanophenol	1	1	0.3	0.749544	1.155988	119.13
4-cyanophenol	1	1	0.2	1.013982	1.155988	119.13
4-cyanophenol	1	1	0.2	1.009698	1.155988	119.13
4-iodophenol	2	1	0.8	-0.16327	2.460667	220.01
4-iodophenol	2	1	0.8	-0.16327	2.460667	220.01
4-iodophenol	2	1	0.8	-0.15748	2.460667	220.01
4-iodophenol	2	1	0.75	0.013788	2.460667	220.01
4-iodophenol	2	1	0.75	0.019566	2.460667	220.01
4-iodophenol	2	1	0.75	0.015723	2.460667	220.01
4-iodophenol	2	1	0.7	0.16869	2.460667	220.01
4-iodophenol	2	1	0.7	0.170045	2.460667	220.01
4-iodophenol	2	1	0.7	0.167331	2.460667	220.01
4-iodophenol	2	1	0.6	0.503646	2.460667	220.01
4-iodophenol	2	1	0.6	0.506773	2.460667	220.01
4-iodophenol	2	1	0.6	0.506773	2.460667	220.01
4-iodophenol	2	1	0.5	0.918088	2.460667	220.01
4-iodophenol	2	1	0.5	0.911514	2.460667	220.01
4-iodophenol	2	1	0.4	1.315787	2.460667	220.01
4-iodophenol	2	1	0.4	1.326015	2.460667	220.01

4. Stan code for "Pooled-log P" model

```
functions{
 real hplcmodel(real fi, real logkw, real logkm, real logS2M, real logka, real logS2A, int modifier){

  real logk; // retention factor
  real S1M; // slope for MeOH
  real S1A; // slope for ACN
  real S1;  // slope
  real lnS2; // curvature coefficient

  S1M = (logkw - logkm)*(1+exp(logS2M)); // curvature coefficient for MeOH
  S1A = (logkw - logka)*(1+exp(logS2A)); // curvature coefficient for ACN
  S1 = S1M * (2-modifier) + S1A * (modifier-1); // combined
  logS2 = logS2M * (2-modifier) + logS2A * (modifier-1);
```

```
 96
 97     logk = logkw - S1 * fi / (1 + exp(logS2) * fi);
 98
 99     return logk;
100      }
101     }
102
103     data{
104     int nAnalytes; // number of analytes
105     int nObs; // number of observations
106     int modifier[nObs]; // modifier (1 for MeOH, 2 for ACN)
107     int analyte[nObs]; // analytes indexes
108      vector[nObs] logkObs; // observed retention factors
109      vector[nObs] fi; // organic modifier content in the mobile phase
110      real MLOGP[nAnalytes]; // molecular descriptor
111
112     intnfiplot;// number of fi for plotting
113      vector[nfiplot] fiplot;// organic modifier content in the mobile phase
114      }
115
116     transformed data{
117      vector[5] etaHat;
118     etaHat= rep_vector(0.0,5);
119      }
120
121     parameters{
122      real logkwHat; // typical logkw for analyte with MLOGP=2.34
123      real logkmHat; // typical logkm for analyte with MLOGP=2.34
124      real logkaHat; // typical logka for analyte with MLOGP=2.34
125      real logS2MHat; // typical curvature coefficient for MeOH
126      real logS2AHat; // typical curvature coefficient for ACN
127      real<lower = 0>sigmaadd; // standard deviation for residuals
128     corr_matrix[5] rho; // correlation matrix
129      vector<lower = 0>[5] omega; // diagonal elements of variance-covariance matrix
130      real beta_logkw; // regression coefficient for logkw
131      real beta_logka; // regression coefficient for logkm
132      real beta_logkm; // regression coefficient for logka
133      real nu; // normality constant
134      vector[5] eta[nAnalytes]; // inter-individual variability
135     }
136
137     transformed parameters{
138     cov_matrix[5] Omega;// variance-covariance matrix
139      real logkw[nAnalytes];// analyte- specific logkw
140      real logkm[nAnalytes];// analyte- specific logkm
141      real logka[nAnalytes];// analyte- specific logka
142      real logS2M[nAnalytes];// analyte-specific curvature coefficient for methanol
143      real logS2A[nAnalytes];// analyte- specific curvature coefficient for acetonitrile
144      vector[nObs] logkHat;// predicted logk
145
146      Omega = quad_form_diag(rho, omega);// diag_matrix(omega) * rho * diag_matrix(omega)
147
148      for(j in 1:nAnalytes){
149     logkm[j] = eta[j, 1] + logkmHat + beta_logkm * (MLOGP[j] - 2.34);
```

```
150   logka[j] = eta[j, 2] + logkaHat + beta_logka * (MLOGP[j] - 2.34);
151     logS2M[j] = eta[j, 3] + logS2MHat;
152     logS2A[j] = eta[j, 4] + logS2AHat;
153     logkw[j] = eta[j, 5] + logkwHat + beta_logkw * (MLOGP[j]- 2.34);
154    }
155    for(i in 1:nObs){
156   logkHat[i] = hplcmodel(fi[i], logkw[analyte[i]], logkm[analyte[i]], logS2M[analyte[i]], logka[analyte[i]], logS2A[analyte[i]],
157   modifier[i]);
158   }
159   }
160
161   model{
162   logkwHat~ normal(2, 5);
163   logkmHat~ normal(0, 5);
164   logkaHat ~ normal(0, 5);
165    logS2MHat~ normal(log(0.2), 0.5);
166    logS2AHat~ normal(log(2), 0.5);
167   beta_logkw  ~ normal(1,0.5);
168   beta_logka  ~ normal(1,0.5);
169   beta_logkm  ~ normal(1,0.5);
170    omega~ normal(0,5);
171   nu ~ gamma(2,0.1);
172    rho~ lkj_corr(1);
173   sigmaadd~ normal(0,1);
174    eta~ multi_student_t(nu,etaHat, Omega);// inter-analyte variability
175   logkObs~ normal(logkHat, sigmaadd);// observations
176   }
177
178   generated quantities{
179    vector[5] etaPred[nAnalytes];// etas
180    real logkwPred[nAnalytes];// retention factor in neat water
181    real logkmPred[nAnalytes];// retention factor in methanol
182    real logkaPred[nAnalytes];// retention factor in acetonitrile
183    real logS2MPred[nAnalytes]; // curvature coefficient for methanol
184    real logS2APred[nAnalytes];// curvature coefficient for acetonitrile
185    vector[nObs] logkHatPred; // predicted logk
186    real logkCond[nObs];
187    real logkPred[nObs];
188    real log_lik[nObs];
189    real log_likPred[nObs];
190    matrix[nAnalytes,nfiplot] logkHatPlotMCond;
191    matrix[nAnalytes,nfiplot] logkHatPlotACond;
192    matrix[nAnalytes,nfiplot] logkPlotMCond;
193    matrix[nAnalytes,nfiplot] logkPlotACond;
194    matrix[nAnalytes,nfiplot] logkHatPlotMPred;
195    matrix[nAnalytes,nfiplot] logkHatPlotAPred;
196    matrix[nAnalytes,nfiplot] logkPlotMPred;
197    matrix[nAnalytes,nfiplot] logkPlotAPred;
198
199    for(j in 1:nAnalytes){
200   etaPred[j] = multi_student_t_rng(nu,etaHat, Omega); // inter-analyte variability
201   logkmPred[j] = etaPred[j, 1] + logkmHat + beta_logkm * MLOGP[j];
202   logkaPred[j] = etaPred[j, 2] + logkaHat + beta_logka * MLOGP[j];
203     logS2MPred[j]= etaPred[j, 3] + logS2MHat;
```

```
 logS2APred[j]= etaPred[j, 4] + logS2AHat;
logkwPred[j]= etaPred[j, 5] + logkwHat + beta_logkw * MLOGP[j];
 }

 for(i in 1:nObs){
logkHatPred[i]= hplcmodel(fi[i], logkwPred[analyte[i]], logkmPred[analyte[i]], logS2MPred[analyte[i]],
logkaPred[analyte[i]], logS2APred[analyte[i]], modifier[i]);
logkCond[i]= normal_rng(logkHat[i], sigmaadd);
logkPred[i]= normal_rng(logkHatPred[i], sigmaadd);
log_lik[i]= normal_lpdf(logkObs[i] | logkHat[i], sigmaadd);
log_likPred[i]= normal_lpdf(logkObs[i] | logkHatPred[i], sigmaadd);
 }

 for(j in 1:nAnalytes){
 for(z in 1:nfiplot){
logkHatPlotMCond[j,z]= hplcmodel(fiplot[z], logkw[j], logkm[j], logS2M[j], logka[j], logS2A[j], 1);
logkHatPlotACond[j,z]= hplcmodel(fiplot[z], logkw[j], logkm[j], logS2M[j], logka[j], logS2A[j], 2);
logkPlotMCond[j,z]= normal_rng(logkHatPlotMCond[j,z], sigmaadd);
logkPlotACond[j,z] = normal_rng(logkHatPlotACond[j,z], sigmaadd);
logkHatPlotMPred[j,z] = hplcmodel(fiplot[z], logkwPred[j], logkmPred[j], logS2MPred[j], logkaPred[j], logS2APred[j], 1);
logkHatPlotAPred[j,z] = hplcmodel(fiplot[z], logkwPred[j], logkmPred[j], logS2MPred[j], logkaPred[j], logS2APred[j], 2);
logkPlotMPred[j,z] = normal_rng(logkHatPlotMPred[j,z], sigmaadd);
logkPlotAPred[j,z] = normal_rng(logkHatPlotAPred[j,z], sigmaadd);
}
 }
}
```

The model was run in Matlab:

```
fileID = fopen('Data.csv','r');
dataArray = textscan(fileID, '%C%f%f%f%f%f%f%[^\n\r]', 'Delimiter', ',', 'TextType',
'string', 'HeaderLines' , 1, 'ReturnOnError', false, 'EndOfLine', '\r\n');
fclose(fileID);

Data = table(dataArray{1:end-1}, 'VariableNames',
{'Name','ID','MeOH1ACN2','fi','logkObs','MLOGP','MMOL'});

fi= 0:0.01:1; % for plots
[uID,iID] = unique(Data.ID);
datastruct                  = struct(...
    'MLOGP',                Data.MLOGP(iID),  ...
    'nAnalytes',            length(uID),...
    'nObs',                 length(Data.ID),...
    'analyte',              Data.ID,...
    'modifier',             Data.MeOH1ACN2,...
    'logkObs',              Data.logkObs,...
    'fi',                   Data.fi, ...
    'nfiplot',              length(fi), ...
    'fiplot',               fi);

fitHPLCizomodel = stan('file', 'HPLCizomodel.stan', 'data', datastruct, 'working_dir',
'tmp', 'iter', 1000, 'warmup', 1000, 'chains', 4, 'init', init0);
```

5. List of abbreviations

1A – predictions scenario based on single measurement in acetonitrile

1M – predictions scenario based on single measurement in methanol

1M & 1A – predictions scenario based on single measurement in methanol and single measurement in acetonitrile

ACN – acetonitrile

AllA – predictions scenario based on all measurements in acetonitrile

AllM – predictions scenario based on all measurements in methanol

AWAS – water–accessible molecular surface area (HyperChem descriptor)

B3LYP – Becke, three–parameter, Lee–Yang–Parr functional

$_i$ (index) – index denoting analyte

$_j$ (index) – index denoting mobile phase composition

$_k$ (index) – index denoting organic modifier type (1 for methanol, 2 for acetonitrile)

LKJ – Lewandowski, Kurowicka, and Joe distribution

log k – retention factor

log k_a – retention factor in the neat acetonitrile

log k_m – retention factor in the neat methanol

log k_{Obs} – observed log k

log k_w – retention factor in the neat water

log P – partition coefficient

LPDCV – Log Predictive Density of Cross Validation

MCMC – Markov Chain Monte Carlo sampling

MeOH – methanol

MLOGP – Moriguchi octanol–water partition coefficient (Dragon descriptor)

MMOL – molar mass (Dragon descriptor)

MST – multivariate student t distribution

N – normal distribution

QSRR – Quantitative Structure–Retention Relationships

RMSECV – Root Mean Squared Error of Cross Validation

RP HPLC – Reversed Phase High Performance Liquid Chromatography

S_1 – slope coefficient in the Snyder–Soczewiński equation

S_2 – curvature coefficient in the Neue *et al.* equation

SMILES – Simplified Molecular Input Line Entry Specification

STO – Slater–Type Orbitals

WAIC – Watanabe–Akaike Information Criterion

$_z$ (index) – index denoting measurement

β – regression coefficient

δ_{min} – maximum electron excess on a most charged atom (HyperChem descriptor)

θ – mean value of the parameter

μ – total dipole moment (HyperChem descriptor)

ν – normality parameter

ρ – correlation matrix in the multivariate student t distribution

σ – standard deviation

φ – organic modifier content

ω – scale parameter in the multivariate student t distribution

Ω – variance–covariance matrix in the multivariate student t distribution