Improved hydrophobic subtraction model of reversed-phase liquid chromatography selectivity based on a large dataset with a focus on isomer selectivity

Sarah C. Rutan[a], Trevor Kempen[b], Tina Dahlseid[b], Zachary Kruger[b], Bob Pirok[b], Jonathan G. Shackman[c], Yiyang Zhou[c], Qinggang Wang[c] and Dwight R. Stoll[b,*]

[a]Department of Chemistry, Box 842006, Virginia Commonwealth University, Richmond, VA 23284-2006, USA

[b]Department of Chemistry, Gustavus Adolphus College, 800 W. College Ave., St. Peter, MN 56082, USA

[c]Chemistry Process Development, Bristol Myers Squibb, 1 Squibb Dr., New Brunswick, NJ 08903, USA

[*]Corresponding author: E-mail address: dstoll@gustavus.edu (D. R. Stoll)

Keywords: hydrophobic subtraction model; isomer selectivity; pharmaceuticals; principal components analysis

ABSTRACT

Reversed-phase (RP) liquid chromatography is an important tool for the characterization of materials and products in the pharmaceutical industry. Method development is still challenging in this application space, particularly when dealing with closely-related compounds. Models of chromatographic selectivity are useful for predicting which columns out of the hundreds that are available are likely to have very similar, or different, selectivity for the application at hand. The hydrophobic subtraction model (HSM1) has been widely employed for this purpose; the column database for this model currently stands at 750 columns. In previous work we explored a refinement of the original HSM1 (HSM2) and found that increasing the size of the dataset used to train the model dramatically reduced the number of gross errors in predictions of selectivity made using the model. In this paper we describe further work in this direction (HSM3), this time based on a much larger dataset (43,329 total measurements) containing selectivities for compounds covering a broader range of physicochemical properties compared to HSM1. This includes multiple compounds that are actual active pharmaceutical ingredients and related synthetic intermediates and impurities, as well as multiple pairs of closely related structures (e.g., geometric and cis-/trans- isomers). The HSM3 model is based on retention measurements for 75 compounds using 13 RP stationary phases and a mobile phase of 40/60 acetonitrile/25 mM ammonium formate buffer at pH 3.2. This data-driven model produced predictions of ln α (chromatographic selectivity using ethylbenzene as the reference compound) with average absolute errors of approximately 0.033, which corresponds to errors in α of about 3 %. In some cases, the prediction of the trans-/cis- selectivities for positional and geometric isomers was relatively accurate, and the driving forces for the observed selectivity could be inferred by examination of the relative magnitudes of the terms in the HSM3 model. For some geometric isomer pairs the interactions mainly responsible for the observed selectivities could not be rationalized due to large uncertainties for particular terms in the model. This suggests that more work is needed in the future to explore other HSM-type models and continue expanding the training dataset in order to continue improving the predictive accuracy of these models.

1. Introduction

Reversed-phase liquid chromatography (RPLC) is an essential tool for the analysis of target analytes in a wide variety of scientific investigations. RPLC has been for years a predominant technology in the pharmaceutical industry for stability indicating methods to establish impurity profiles for drug substances, drug products, intermediates and in-process control samples. However, it is currently challenging to select appropriate LC method conditions (i.e., stationary phases and mobile phases) for a target separation without time-consuming method development studies.

In order to support method development efforts, it is useful to have models for chromatographic selectivity that are global in scope, such that the model can accommodate both charged and neutral molecules, large and small molecules, and a diversity of stationary phases. At the same time, it is desirable to have models that can accurately predict the selectivity for the separation of highly similar molecules, especially isomeric compounds. These latter separations can be particularly challenging, but they are critically important in contemporary pharmaceutical analysis.

Quantitative structure retention relationships (QSRRs) have been used for help in the prediction of retention parameters to reduce method development times [1–3]. These models establish a relationship between a chromatographic retention parameter and a set of physiochemically relevant molecular descriptors. Some descriptors can be obtained experimentally, such as octanol-water coefficients (log *P*) [4] and Abraham solute descriptors [5–8], but often these descriptors are obtained from computational molecular geometry optimizations [1]. Some of the most successful models are obtained when groups of structurally similar compounds are considered and local models are developed, because a global, mechanistic model for liquid chromatography has not yet been developed [9].

The hydrophobic subtraction model (HSM; hereafter, HSM1) for RPLC has been in use for over 20 years now [10–20]. This model can be considered a ‘data-driven’ model, in that the solute and stationary phase parameters are derived from retention measurements, rather than externally calculated or measured physicochemical parameters. The HSM1 provides descriptive parameters for RPLC stationary phases that relate to their hydrophobicities, hydrogen bonding capacities, capacities for involvement in ionic interactions, and the contributions of steric effects to their overall selectivities. These characteristics are obtained from the following equation.

$$\log_{10}\alpha = \log_{10}\left(\frac{k_x}{k_{EB}}\right) = \eta' H - \sigma' S^* + \beta' A + \alpha' B + \kappa' C \qquad (1)$$

where α is the chromatographic selectivity for a selected solute, *x*, relative to ethylbenzene (EB) and η’, σ’, β’, α’ and κ’ are solute specific parameters for the solute hydrophobicity, steric effects, hydrogen bond basicity, hydrogen bond acidity, and cation exchange propensity, respectively. The *H*, *S**, *A*, *B* and C parameters are the corresponding descriptors for the stationary phases relevant to specific mobile phase conditions (50/50 acetonitrile (ACN)/60 mM potassium phosphate buffer at pH 2.8). The original model was developed using a set of retention data for 67 solutes on ten type B silica phases [10,11], with an additional 20 solutes added soon afterwards [12]. Subsequent work identified a subset of 15 solutes to be used as probe solutes [13] for routine characterization of stationary phases in different laboratories. To establish the initial HSM1 database, retention factors for these probes, along with ethylbenzene as the reference solute, were determined for a total of 87 RPLC columns (mostly alkyl phases) [13]. Since the early 2000’s, these solutes have been used to establish column parameters for about 750 RPLC stationary phases [21,22].

While the HSM1 has been used widely, it has been recognized that it is not really a global model. A small number of relatively simple molecules has been chosen for routine stationary phase characterization, and the initial model was developed based on using stationary phase chemistries of relatively limited scope (i.e., mainly alkyl phases). Furthermore, we have shown that the model does not carry the information needed to rationalize changes in the selectivity of cis/trans isomers in response to changes in the properties of a RPLC column [23].

Recently, some of us have reevaluated the original dataset as a whole (15 solutes × ~700 stationary phases), to determine whether or not the HSM1 could be refined to reveal more information about RPLC selectivity, since the original model was based on a relatively small number of stationary phases [24]. A revised model, HSM2, based on six parameters, was proposed which takes the following form

$$\log_{10}\alpha = \log_{10}\left(\frac{k_x}{k_{EB}}\right) = hH + bA + aB + kC + vV + dD \qquad (2)$$

Here, *h*, *b*, *a*, *k*, *v*, and *d* are solute parameters for hydrophobicity, hydrogen bond basicity, hydrogen bond acidity, cation exchange propensity, size and dipolarity, respectively, and *H*, *A*, *B*, *C*, *V* and *D* are the complementary stationary phase parameters. Both the original HSM1 and HSM2 are 'data-driven' models, in that the actual retention data are used to make the parameter scales. In the case of HSM1, an iterative subtraction method was used to determine the scales, while for HSM2, principal components analysis (PCA) was used to find scales that were consistent with the selectivity data. While HSM2 was based on a large, relatively diverse set of stationary phases, the 15 solutes used to generate the model were small molecules (i.e., molecular weights were all less than 280 Da) with a somewhat limited hydrophobicity range (log *P* ranging from -0.9 to 4.4) that cannot be considered as representative of the range of solutes that can be analyzed by LC methods, especially compounds of pharmaceutical interest.

We concluded that HSM2 had a chance of better reflecting the chemical richness present in the 750 stationary phases that comprise the current HSM1 database, which include a much broader range of chemistries than the alkyl phases that were used to parameterize the original HSM1 [24]. However, we were still limited to the 15 original solutes, which we were convinced did not capture the broadest range of solute behavior – these molecules are quite simple. Molecules encountered in pharmaceutical analysis exhibit a large range of polarity and molecular weight, and often closely related compounds and isomer pairs must be separated during the drug development process. An example of a situation where the cis/trans selectivity could not be predicted or rationalized is a recent study on the effect of column aging on the cis/trans selectivity of a Bristol Myers Squibb compound, denoted as BMS-A (denoted as Lin-A in this paper). It was found that HSM1 was not able to help predict or rationalize the changes in the cis/trans selectivity for this compound upon column aging [23].

Therefore, in the present study, we have attempted to address the primary limitations of the previous studies: 1) the HSM1 dataset is composed of retention measurements made with just one mobile phase composition (50/50 ACN/buffer), which precludes any direct application of the model to gradient elution conditions; 2) the buffer used for HSM1 contains phosphates, which are incompatible with mass spectrometric detection – an essential tool in the analysis of pharmaceuticals; and 3) the probe solutes have been limited to a small number of relatively simple compounds. In this work, we have produced a large set of retention measurements using our high-throughput method for characterizing retention described previously [25–27]. The new dataset includes 86 solutes and 13 stationary phases, and retention has been measured at multiple mobile phase compositions for each compound/column combination, for a total of about 40,000 measurements. The 13 phases were chosen to cover a broader range of the reversed-phase chemistry reflected in the HSM1 database. The solutes were chosen to include many of the important probes used in other selectivity tests for RPLC (e.g., Tanaka, Engelhardt, etc.; see Table 3 of refs. [28,29]), and also include several compounds of pharmaceutical importance, including positional isomers, and isomer pairs with shape variations. The set also includes molecules with molecular weights of up to 600 Da, and the logP values range from 0.2 to 6.0. The 13 stationary phases were selected from the larger set of stationary phases used in the development of HSM2, with an eye towards the selection of phases with the widest differences in selectivity, as well as phases of practical use in the pharmaceutical industry. In this work, we describe the analysis of this dataset that results in a new HSM-type model (HSM3), with a focus on determining whether we could achieve improvement in the prediction of isomer selectivities.

2. Materials and methods

2.1 Data collection

Retention factors were determined for 89 solutes on 13 stationary phases using mobile phases composed of ACN and an aqueous buffer containing ammonium formate (25 mM in ammonium and 105 mM in formate) at pH 3.2. The LC instrument was composed of modules from Agilent Technologies (Waldbronn Germany): binary pump (G4220A), autosampler (G7167B), thermostatted column compartment (G7116B), and diode array UV absorbance detector (G4212A). As described in ref. [26], samples were introduced to the mobile phase stream using a “feed injection” approach, and the injection volume was 150 nL. The solutes and stationary phases used are listed in the supplementary materials in Tables S1 and S2. Our high-throughput measurement approach is based on retention measurements made using very short columns (typically 5 to 20 mm in length and 2 mm in diameter), and then corrected using the retention factor of toluene measured using a conventionally sized column (typically 100 mm x 2.1 mm i.d.). The dimensions of all these columns are given in Table S2. The details associated with the measurement steps and implementation of correction factors were described previously [25,26].

Generally, five replicate retention measurements were made for each solute/stationary phase/mobile phase combination, and mobile phase compositions were chosen so that: 1) retention data were obtained at five different compositions for each solute/column combination; and 2) the lowest retention factor is between 0.5 and 3.0, the highest retention factor is between 15 and 50, and the other three points are roughly evenly spaced between retention factors of 3 and 15. Meeting these criteria was not always possible, for example in the case of highly hydrophilic compounds. When working with a particular column, a set of quality control (QC) measurements were made to enable monitoring of column (e.g., stationary phase aging and column-to-column variability) and system changes over time. Such measurements were made using uracil, toluene, ethylbenzene, 4-n-butylbenzoic acid, 4-n-hexylaniline, and nortriptyline as QC solutes. Generally, QC measurements were made about once per day. While the entire dataset is composed of measurements made using multiple mobile phase compositions, the model development primarily involves the use of data from a 40/60 ACN/buffer mobile phase. A more thorough exploration of the entire dataset set is left for future work.

The sources of the test solutes are shown in Table S1. Stock solutions were prepared at 10 mg/mL, typically in ACN, or 50/50 ACN/water if they were not soluble in ACN. Then, a working solution was prepared at either 0.2 or 5.0 mg/mL in either ACN or 50/50 ACN/water.

The full retention dataset used in this work (43,329 measurements) is provided as Supplemental Information in the file "WC_second_kernel_database.xlsx", along several files containing quality control (QC) data as outlined in the Supplemental Information. Note that a subset of the full dataset shared here was published previously (12,319 measurements) [26], and we provide them again here simply for the convenience of the reader.

2.2 Parameter estimates

Calculated parameters for each of the examined solutes were obtained from several sources. Octanol/water partition coefficients (*P*), Connolly solvent-excluded volumes (*V*), molar refraction (*MR*) and ovality (*O*) parameters were calculated using Chem3D (Revvity Signals, v. 20.1.1.125) after MM2 geometry optimization. The shortest dimension of each solute molecule was calculated from the volume and ovality by assuming an oblate spheroid shape. Linear solvation energy relationship (LSER) parameters [5] were obtained from the LSER2017 calculation engine [30]. These parameters included the dipolarity-polarizability (*S*), the polarizability (*E*), the hydrogen bond acidity (*A*) and hydrogen bond basicity (*B*). Acid/base ionization constants for the ionizable solutes were calculated using ACD/Percepta Ver. 2022.2.3 (Advanced Chemistry Development, Inc., Toronto, ON, CA).

2.3 Data analysis

As is discussed in Section 3.1, the HSM3 model was developed using retention factors determined in a mobile phase of 40/60 ACN/buffer. However, experimental measurements were not feasible in this mobile phase for all solute/column combinations because they were impractically large (i.e., > 50). In those cases, the experimental retention factor data we did have were fit to the Neue Kuss (NK) model describing the retention as a function of the volume fraction of organic solvent in the mobile phase (ϕ).

$$k = k_w (1 + S_2\phi)^2 \exp\left(-\frac{S_1\phi}{1+S_2\phi}\right) \tag{3}$$

where k_w, S_1 and S_2 are solute/condition-specific model parameters. The fitting was carried out using a re-parameterization of the NK model where the model parameters were calculated based on the retention factor at $\phi = 0.30$ as a reference point (k_{ref}) instead of the more conventional k_w, as described in a recent publication [31]. The model is then given in revised form as

$$k = k_{ref}\left(1 + S_{2,ref}\left(\phi - \phi_{ref}\right)\right)^2 \exp\left(-\frac{S_{1,ref}\left(\phi - \phi_{ref}\right)}{1 + S_{2,ref}\left(\phi - \phi_{ref}\right)}\right) \tag{4}$$

Here, ϕ_{ref} is taken as 0.30, and k_{ref}, $S_{1,ref}$ and $S_{2,ref}$ are the re-parameterized model parameters. Fits to this equation were carried out using the *fitlm* function in the Statistics and Machine Learning Toolbox in Matlab (Mathworks, Natick, MA).

All other data analyses were carried out in Microsoft Excel and using standard functions in Matlab.

3. Development of model

3.1 Initial construction of dataset

The original HSM1 model and HSM2 were based on retention measurements made using 50/50 ACN/60 mM potassium phosphate at pH 2.8. In this work we have elected to focus on data obtained using a mobile phase containing 40% ACN, because many compounds that we think are important to model are simply not retained well enough in 50% ACN to use the data reliably. Also, our high throughput retention measurement approach makes it more feasible to measure retention factors up to 50 than in the past when the use of 150 mm x 4.6 mm i.d. columns was the norm. Additionally, we elected to use a more 'mass spectrometry friendly' buffer of ammonium formate. The complete dataset of retention values in 40/60 ACN/buffer consists of 89 x 13 = 1157 retention factors (this corresponds to 86 unique compounds, because of three duplicate measurements.). However, in 72 of the 1157 cases, the retention factor at 40% ACN was not measured experimentally, in most cases because the retention factor was too large to be practically determined at this mobile phase composition. Therefore, these missing values were estimated by fitting the available data for those column/solute combinations to the NK model as described above [31]. This methodology allowed for the rejection of outliers [31], and provided stable estimates for the NK parameters. For three solutes – 2,2'-dinaphthyl ether, glecaprevir and o-terphenyl – more than 50% of the retention factors on the 13 columns were missing, because of very high retention, and these solutes were eliminated from further analysis. Furthermore, eight additional solutes showed very low retention on some of the columns. These solutes are (with the median retention factors for the 13 columns shown in parentheses) 2-nitrobenzoic acid (0.32), 4,4'-dipyridyl (0.31), benzyltrimethylammonium chloride (0.14), caffeine (0.26), dasatinib (0.61), N-benzylformamide (0.62), pyridine (0.15) and risperidone (0.70). These low retention factors lead to very high standard deviations in ln α of 1.5 to 31. Because the PCA analysis and subsequent linear regression modeling are based on the data having similar variances, we elected to remove these solutes from the dataset as well. The distribution of the remaining 78 x 13 = 1014 retention factors (in terms of ln *k*) is shown in Fig. 1A and the corresponding box and whisker plot is shown in Fig. 1B. Fig. 1B also shows the box and whisker plot for the pharmaceutical compounds only. Similar plots are shown for the distribution of the ln α values in Figs. 1C and 1D. The values in red in Figs. 1A and 1C are those values estimated from the NK model. The mean standard deviation of the ln α values is 0.0528 and the median standard deviation is 0.0174. The final 78 solutes are shown in the supplemental material in Table S1, and the 13 selected stationary phases are shown in Table S2.

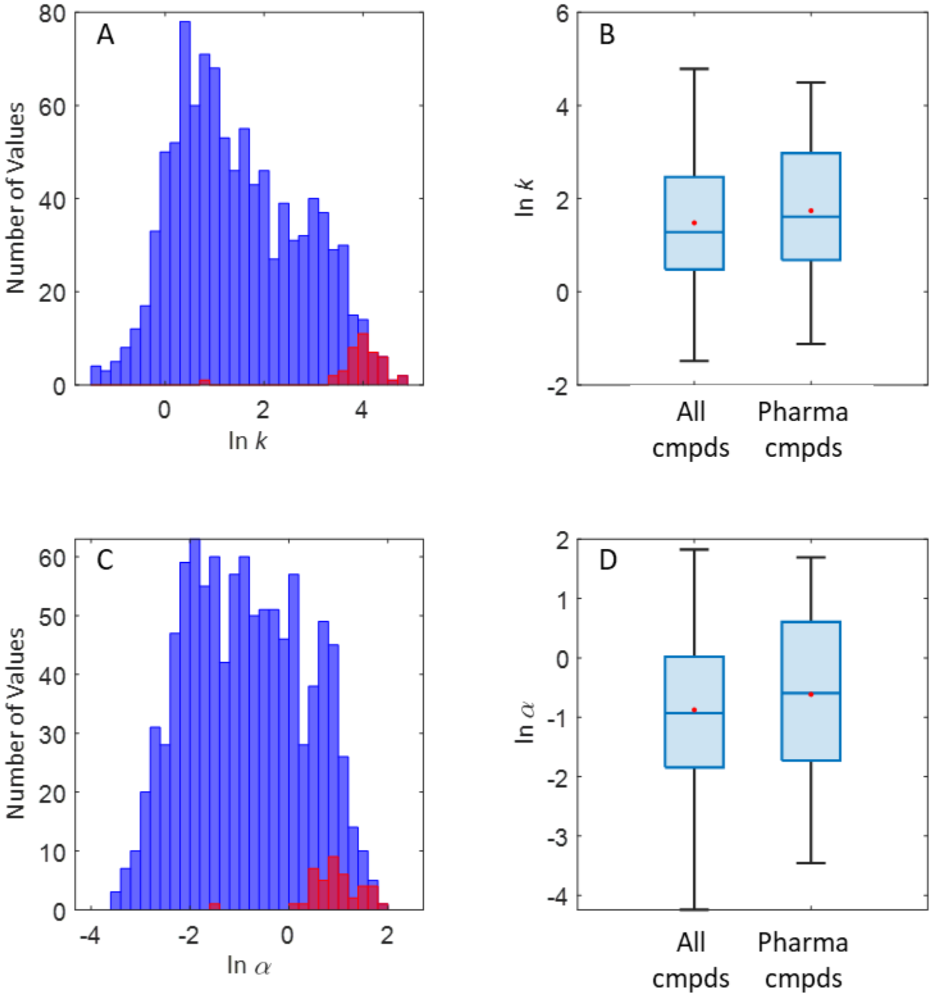


Figure 1. (A) Histogram of ln k values in entire 78x13 data set. Values in red indicate those values that were estimated from the NK equation. (B) Box plot for ln k for all compounds, and for just the pharmaceutical compounds. Whiskers indicate the data range and the boxes indicate the interquartile range. The center line is the median and the dot is the mean. (C) Histogram of ln α values in entire 78x13 data set. (D) Whiskers indicate the data range and the boxes indicate the interquartile range. The center line is the median and the dot is the mean.

This dataset now contains several compounds of interest to the pharmaceutical industry, including some common active pharmaceutical ingredients (APIs) and a set of process impurities and geometric isomers for the API Linrodostat [23,32]. The structures of these pharmaceutical compounds are shown in Fig. 2. The original 15 solute HSM1 dataset did include four pharmaceutical compounds, denoted by the boxes in Figure 2. It can be seen that the structural variability of these compounds is much greater than in the original data set. The physicochemical properties are also highly variable, and several of these properties are given in Table S1.

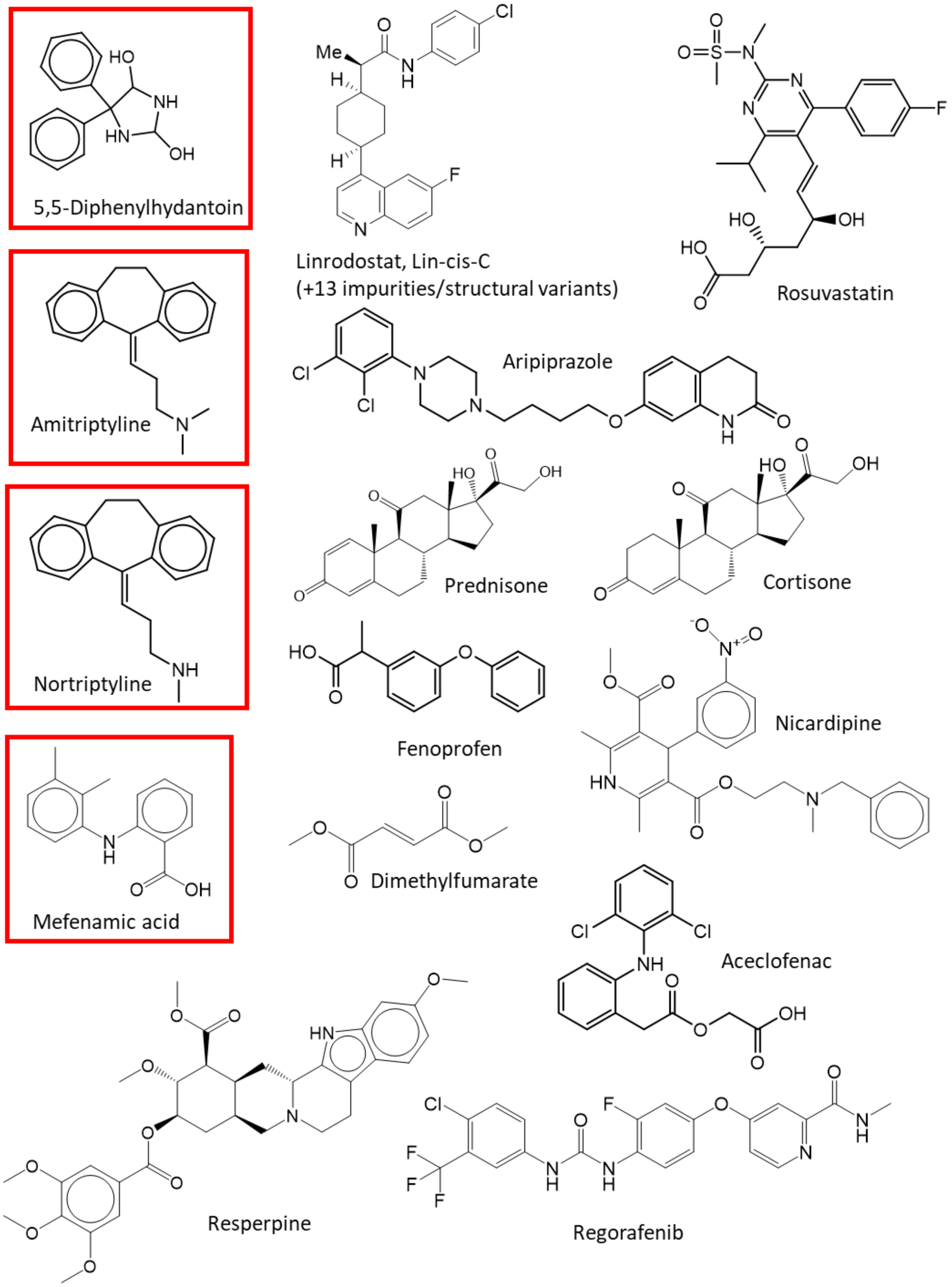
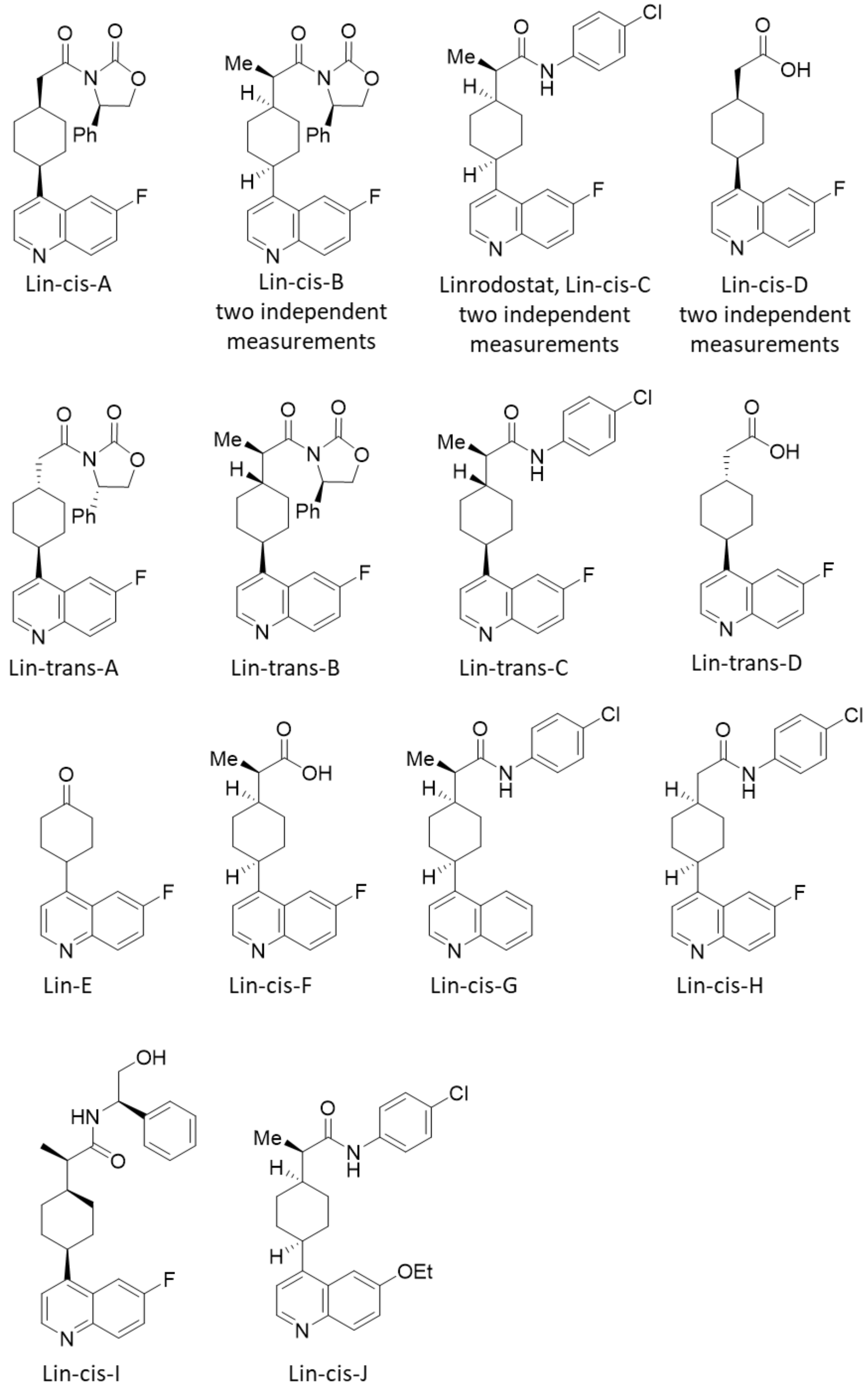


Figure 2. Pharmaceutical compounds in dataset. Compounds boxed in red were in the original HSM1 dataset.

While it is useful to have models that can accommodate compounds with a wide range of physicochemical properties, in the pharmaceutical industry it is often the case that the API must be resolved and analyzed in mixtures containing many similar compounds (e.g., starting materials, intermediates, process impurities and degradants). To this end, the data set also includes a number of compounds of this nature that are related to the API Linrodostat. The structures of these compounds are shown in Fig. 3. Most of these compounds contain a core (6-fluoroquinolin-4-yl)cylclohexyl structure, giving them a moderate to high degree of structural similarity. The inclusion of these compounds in the dataset allowed us to evaluate whether the model can lead to insights into the chromatographic selectivity for the types of closely related compounds that need to be resolved and analyzed in pharmaceutical drug development research.



Lin-cis-A

Lin-cis-B two independent measurements

Linrodostat, Lin-cis-C two independent measurements

Lin-cis-D two independent measurements

Lin-trans-A

Lin-trans-B

Lin-trans-C

Lin-trans-D

Lin-E

Lin-cis-F

Lin-cis-G

Lin-cis-H

Lin-cis-I

Lin-cis-J

Figure 3. Structures of linrodostat and related compounds.

3.2 Development of parameter scales

An initial PCA of the 78 x 13 dataset indicated that 6-7 PCs could be justified, based on a minimum in the root-mean-square error of cross validation (RMSECV) (via leave-one-out cross validation). The RMSECV for 6 PCs was 0.383, and the RMSECV for 7 PCs was slightly higher, at 0.389. Note that these errors are significantly higher than the root-mean-square error of calibration (RMSEC), which were 0.0390 and 0.0330, for 6 and 7 PCs respectively. This is because at least one of the 13 columns (Bonus RP) exhibited unique selectivity relative to the other 12 columns.

To better evaluate the performance of the PCA model, we elected to split the data into training and validation sets. Several methods have been proposed for the selection of training and validation sets [33,34]. On one hand, the training set should be representative of the variability in the original data set, but if this leaves only compounds in the validation set that are highly similar to the training set, the validation set metrics will be too optimistic. Alternatively, the selection of the training set and validation sets can be done completely randomly, but the process must be repeated multiple times, because some of the training sets chosen will inevitably not sample the whole model space. The solutes were allocated to the two sets to make sure that molecules with the same general structural features were included in both the training and validation sets. The training set contained a little more than twice as many compounds as the validation set (56 compared to 22); these sets are denoted in Table S1. We first focused our attention on the 56 solute training set. A plot of the first two PCs for this data set is shown in Fig. 4; the solutes corresponding to the numbered points are given in Table S1. The general trends in this plot are interesting – the points that bracket the sloping group of points at the bottom of the figure correspond to N,N-dimethylbenzamide (39, pink) and triphenylene (56, blue-green), a relatively hydrophilic and a relatively hydrophobic compound, respectively. The log *P* for N,N-dimethylbenzamide is 0.62 and the log *P* for triphenylene is 5.23. The points clustered at the top left of the figure (shown in blue) correspond to amitriptyline (18), aripiprazole (20), berberine (22), nicardipine (42), nortriptyline (44) and reserpine (52). These are all ionized or ionizable bases. Interestingly, three points deviate below the hydrophobic trend line, 2,4-dinitrophenol (9), 4-n-butylbenzoic acid (11) and mefenamic acid (38) (shown in red) and have pK_as of 4.2, 4.1 and 4.3, respectively. These acidic solutes are likely partially ionized under these separation conditions (although it is difficult to quantify the effect of acetonitrile on the degree of ionization). Therefore, the first PC approximately correlates with hydrophobicity, while the second PC approximately correlates with the likelihood of a solute interacting with the stationary phase via ionic interactions. This is consistent with the development of the original HSM1 which found that the primary and secondary contributions to the selectivity were hydrophobicity and ionic interactions, respectively. The RMSEC values for the training and validation sets for 6, 7 and 8 PCs are shown in Table 1 [10]. An F-test shows that the validation set RMSEC is not significantly greater than the training set RMSEC for the 7 PC model, while the validation set RMSEC is significantly greater than the training set RMSEC for the 8 PC model. Thus, we proceeded with model development using a 7-component model. A plot of the predicted ln α vs. the actual ln α values is shown in Fig. 5A, and the residuals are shown in Fig. 5B, with the training set points represented by the red circles, and the validation set points represented by the blue squares.

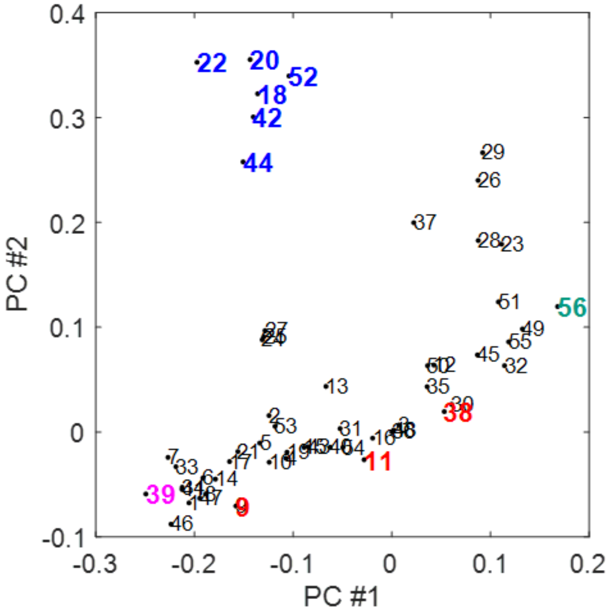


Figure 4. Plot of the first 2 PC’s for the 56 x 13 training set ln α dataset. Point 39 is N,N-dimethylbenzamide (pink), point 56 is triphenylene (blue-green), points 18, 20, 22, 42, 44, and 52, amitriptyline, aripiprazole, berberine, nicardipine, nortriptyline, and reserpine, respectively (blue), and points 9, 11, and 38, 2,4-dinitrophenol, 4-n-butylbenzoic acid and mefenamic acid, respectively (red). See Figure S1 for the number correspondence for the other solutes.

Table 1. RMSEC values for training and validation sets.

Model	Training	Validation	F^a
6 PCs	0.0396	0.0389	0.965
7 PCs[b]	0.0333	0.0339	1.036
8 PCs	0.0264	0.0298	1.274
Raw parameters	0.627	---	---

[a] F_{crit} = 1.173 (p = 0.05). [b] These are the RMSEC values for the final HSM3 model as well.

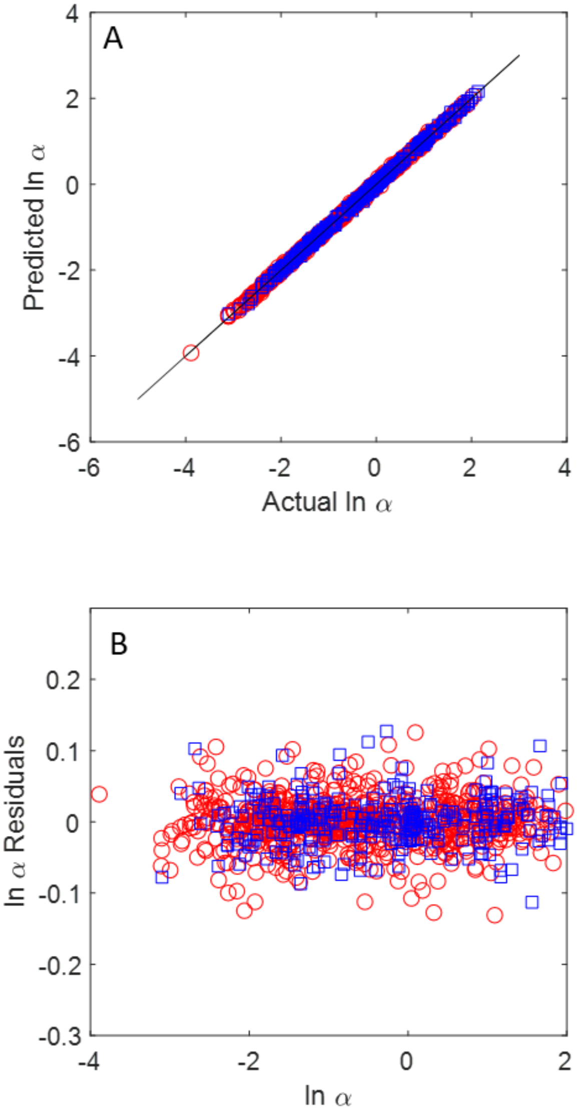


Figure 5. (A) Predicted ln α from 7 PC model vs. the actual ln α. (B) ln α residuals. Red circles are the training data, and blue squares are the validation data.

While we could use this PCA model for prediction of ln α, we wanted to find a model that provided some chemical rationale for the observed selectivities. Although the first two PCs were found to roughly correspond with hydrophobicity and ionic interactions, respectively, the remaining PCs showed no obvious correlations with known chemical behavior. We wanted to find directions in the 7-dimensional PC space that better represented known chemical behavior, while still relying on a data-driven model to have the best predictive accuracy. However, we wished to avoid using more chemically relevant parameters at the expense of the model stability. A PCA model is inherently the most stable model, in that there are no collinearities between the PCA axes, by definition. Mathematically, this corresponds to the solute PC matrix having a condition number of 1. Any model other than the PCA model will have a condition number greater than one. Models with high condition numbers will not allow for precise parameters to be calculated for new stationary phases/solutes.

We evaluated several candidate solute parameter scales as targets to 'rotate' the PC axes toward more chemically interpretable parameters. The final candidate scales chosen are shown in Table 2. Each of these parameter scales was fit to a linear regression model of the 7 PCs. The resulting fitted predictions were used to form each of the corresponding solute parameter scales. Note that we also considered using robust linear regression (used in the HSM2 model development) for this step [24] as opposed to classical linear regression, but there were only minor differences in the outcomes from the two approaches, so classical regression was used. The training and validation RMSEC values for the final parameter scales initiated from those shown in Table 2 fit to the ln α values were identical to the values shown in Table 1 (0.0333 and 0.0339 for the training and validation sets, respectively) for the 7 PC model, because the final parameter scales are simply a rotation of the PC values. The parameter values for all 78 solutes and for all 13 stationary phases are shown in the Excel spreadsheet provided in the Supplemental Information, as well as Tables S3 and S4. The final model is therefore given as

$$\ln\alpha = \ln\left(\frac{k_x}{k_{EB}}\right) = hH + kC + aB + bA + dD + eE + sS \qquad (5)$$

Table 2. Final target parameter scales

Target Scale	**Source**	**Physicochemical Effect**	**Solute Parameter**	**r^2 (7 PCs) Training Set**	**r^2 Final Model**
Log P_{cd}/2	Chem3D[a]	Hydrophobicity	*h*	0.7902	0.8321
$(\alpha_+ - 30\ \alpha_-)MR/100$	ACD/Labs[b] (α values from pK_as); *MR*[c] (Chem3D)	Ionic interactions	*k*	0.6674	0.6831
E	LSER 2017 calculation[d]	Polarizability	*e*	0.8063	0.8080
$S - mE - b$	LSER 2017 calculation[d,e]	Dipolarity	*d*	0.6123	0.2580
A	LSER 2017 calculation[f]	Hydrogen bond acidity	*a*	0.4964	0.5416
B	LSER 2017 calculation[g]	Hydrogen bond basicity	*b*	0.6078	0.6633
Oblate spheroid minor axis, truncated so that values <0.04 are set to zero	Ovality and *V* (Chem3D)[h]	Steric exclusion	*s*	0.6123	0.6915

[a] log *P* of octanol water partition coefficient calculated in Chem3D (Revvity Signals, v. 20.1.1.125); [b] pK_a values of ionizable acids and bases from ACD/Labs ACD/Percepta Ver. 2022.2.3 (Advanced Chemistry Development, Inc., Toronto, ON, CA), $\alpha_+ = [H^+]/([H^+] + K_a)$, $\alpha_+ = K_a/([H^+] + K_a)$; [c] molar refraction calculated in Chem3D (Revvity Signals, v. 20.1.1.125); [d] LSER polarizability (*E*) calculated from LSER 2017 [30]; [e] LSER dipolarity/polarizability (*S*) calculated from LSER 2017 [30]; [f] LSER hydrogen bond acidity (*A*) calculated from LSER 2017 [30]; [g] LSER hydrogen bond basicity (*B*) calculated from LSER 2017 [30]; [g] Dimension of the minor axis assuming an oblate spheroid shape based on ovality and Connolly solvent-excluded volumes calculated in Chem3D (Revvity Signals, v. 20.1.1.125).

As we explored different scales and different combinations of scales, we sought to find final parameter scales with a condition number as close to one as possible. During this process, we found condition numbers as high as 200-300. The condition number for the final solute parameter matrix expressed by Eq. (5) is 16.8. This is a satisfactory result, especially because by their very nature, we expected some degree of correlation in the various solute parameter scales.

It is instructive to pause and examine the correlation between the initial target parameter scales and the final parameter scales obtained from fitting to the PCs. The correlations for the original parameter scales to the parameter scales for the training set and for the final model parameters are shown in Table 2. None of the correlations are particularly strong. This lack of correlation indicates that the original scales do not entirely capture the physicochemical properties revealed from the data-driven model. Additionally, the average standard error of the training data set fits based on the raw parameter scales was 0.627 – this is more than 15-fold worse than the RMSEC values for the PCA model and the final model shown in Table 1. The likely reasons for this much larger error are (1) that the 'true' model may not be a linearly additive model, as assumed here, (2) that even for those scales that are derived from measured parameters (e.g., log *P*) the parameters are derived from a different physicochemical partitioning process (i.e., different pHs, solvents and temperatures), (3) that some of the parameters are based on structures optimized in the gas phase (e.g., volume, ovality and K_a), and (4) that many of the parameters are estimated from linear regression models themselves (e.g., the LSER parameters). It is clear that simply using pre-established physicochemical parameters does not produce an adequate model, whereas the data-driven model gives very promising results.

From the signs of the solute and column parameters, we can make some generalizations as to the effects of the physicochemical properties on retention (at least for the subset of columns studied here). Solutes with larger hydrophobicity (*h*), that are more polarizable (*e*), that are hydrogen bases (*b*) and that are larger molecules (*s*) all will be retained more strongly on these stationary phases (the column parameters for these properties are all positive, see Table S4, except for negative values for *A* for the Bonus RP and Eclipse PAH phases). The increase in retention with increasing size was not what we expected, as we thought that this term might reflect lesser retention for the largest molecules because of steric exclusion from the stationary phase [35–38]. However, this parameter does show differences between the sizes of the cis- and trans- geometric isomers, which reflects what can be seen in the 3D representation of these molecules. Visually, the cis-structures of the Lin-A, Lin-B, Lin-C and Lin-D compounds appear to have a more compact structure than the corresponding trans isomers, and the *s* parameters for the cis structures are smaller than those for the trans isomers. There are only minute differences in the Connolly solvent-excluded volumes of the isomers calculated by Chem3D, so this parameter would not help in distinguishing the size differences that are captured by the *s* parameter. More dipolar molecules will be retained less, as indicated by the negative *D* parameter coupled with positive *d* values for molecules that are more dipolar than ethylbenzene. The effect of solute hydrogen bond acidity is mixed – on some columns hydrogen bond acids are more retained and on others, less. This latter effect may be due in part to this parameter being mixed with other unidentified physicochemical effects.

Negatively charged molecules (negative *k* values with negative *C* values) are also slightly more retained. These molecules are ionizable acids, such as 2,4-dinitrophenol, 4-n-butylbenzoic acid and mefenamic acid, as mentioned above. In contrast, positively charged species (positive *k* values with negative *C* values) are retained less. This implies that at this pH (3.2) and mobile phase conditions, the stationary phase has a positive charge. While it is well-known that at higher pHs the residual silanols will have a negative charge the possibility of the surface having a positive charge at lower pHs has not been widely recognized [39,40]. Neue et al. noted that one positively charged analyte (the bretylium ion), eluted before the dead volume marker on XTerra RP18 stationary phases [39]. Additionally, Méndez et al. reported anion exchange-based retention based on the retention of the nitrate anion at lower pHs on a Symmetry C18 phase [41].

The overall prediction of the ln α values from the present model (Eq. (5)) vs. HSM1 and HSM2 can be compared by regression of the HSM1 and HSM2 column parameters to the experimental ln α values used in this study. These predictions are shown in Fig. 6. The corresponding standard errors for HSM1, HSM2, and the present model are 0.134, 0.158 and 0.0337, respectively. Interestingly, the HSM2 predictions are not as good as those of the original HSM1 model. Note also that no correction has been made for the fact that the HSM1 and HSM2 column parameters are based on retention measurements where the aqueous buffer was pH 2.8 and 50% ACN, but the retention measurements described here were obtained at pH 3.2 and 40% ACN. Because of these differences in pH and mobile phase composition, this is not an entirely fair comparison. Some of the largest residuals from the HSM1 and HSM2 models are for compounds with larger *k* parameters, as expected because of the difference in pH; this can be seen in Figures 6A and 6B, where those solutes with larger *k* parameters are shown in red.

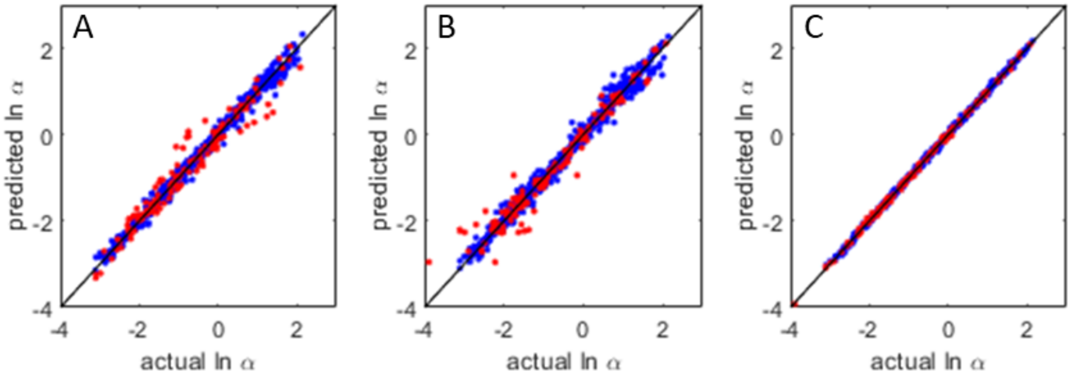


Figure 6. Predicted ln α values vs. actual ln α for (A) HSM1, $s_E = 0.135$; (B) HSM2 , $s_E = 0.158$; (C) current model , $s_E = 0.0337$. Red points are for solutes with $|k| > 0.2$.

The solute parameters determined here are fully 'data-driven' parameters, in that the model expressed by Eq. (5) has the same predictive capability as the 7 PC models. However, the rotation carried out by regression of the PCs to the selected raw parameter scales should provide parameters that are more consistent with chemical intuition and are at least approximately correlated with the physicochemical parameters used to develop the model. The last column of Table 2 shows the correlation of the final model parameters with the physicochemical scales used to initiate the model. The strongest correlation is the h parameter with the logP value, at 0.83, therefore it is fair to conclude that the h parameter represents the hydrophobicity of the solutes. It is noteworthy that the solute (h) and column (H) parameters are not particularly well correlated with the HSM1 η' and H parameters (data not shown). This is not particularly surprising, as the HSM1 η' is based on the retention of solutes on the SB-C18 column, whereas the HSM3 h parameter is initialized based on log P. 'Hydrophobicity' is inherently a mix of multiple physicochemical interactions, and it is expected that the two scales could have a fundamentally different mix of these interactions. The polarizability parameter e is correlated with the LSER E at 0.81. In contrast, the dipolarity parameter d is not well correlated with the initiating scale, which was the LSER S (dipolarity/polarizability) corrected for the polarizability (LSER E), in an attempt to remove polarizability contributions from the scale. Interestingly, the parameter d is more strongly correlated with the original LSER S parameter, at 0.53 (data not shown). We are not too surprised that these correlations are not stronger, because these scales either are calculated from gas-phase structures that do not represent condensed phase properties, or are from computer-generated parameters secondary to actual measured properties, as discussed above.

Figures S1-S7 in the Supplemental Information provide the structures and parameters for each of the solutes with the largest and smallest values in the corresponding parameter scale. In general, it can be seen for most parameters there is a reasonable correlation between the structure and the resulting parameter value, at least from chemical intuition.

Within the 78 x 13 dataset we also have three sets of duplicates. These duplicates were from different lots of the same compounds that were measured independently during dataset collection. These compounds are Linrodostat (labeled Linrodostat 1 and Linrodostat 2, compounds 23 and 69 in Table S3), Lin-cis-B (Lin-cis-B 1 and Lin-cis-B 2, compounds 62 and 63) and Lin-cis-D (Lin-cis-D 1 and Lin-cis-D 2, compounds 24 and 25). (Structures of these compounds are shown in Fig. 3, and compound numbers are shown in Table S3.) These duplicates allowed us to evaluate the reproducibility of the resulting parameters. The values for the parameters for these duplicates are shown in Table 3. The agreement in the parameters for the duplicates are all better than 5 %, calculated relative to the range of each parameter scale.

Table 3. Parameter values for duplicates[a]

	Lin-cis-D	**Lin-cis-B**	**Linrodostat (Lin-cis-C)**
h	0/0.002 (0.39 %)	0.240/0.233 (1.5 %)	0.247/0.224 (4.8 %)
k	0.143/0.126 (3.0 %)	0.172/0.156 (3.0 %)	0.137/0.121 (2.9 %)
a	0.159/0.169 (1.6 %)	0.102/0.107 (1.2 %)	0.149/0.155 (1.4 %)
b	0.159/0.169 (2.6 %)	0.217/0.224 (1.6 %)	0.120/0.128 (2.0 %)
d	0.211/0.220 (2.7 %)	0.130/0.134 (1.2 %)	0.082/0.085 (0.90 %)
e	0.133/0.148 (4.2 %)	0.201/0.216 (4.1 %)	0.182/0.195 (3.6 %)
s	0.212/0.214 (0.4 %)	0.384/0.378 (1.0 %)	0.257/0.259 (0.36 %)

[a] Value in parenthesis corresponds to the % difference between the duplicates relative to the full range of the parameter scale.

The results of the regression analysis also permit the evaluation of the precision of the column and solute parameters. For the column parameters, the percent relative errors in each parameter for each column were calculated from the standard errors of the parameters. The average percent relative error was less than 5% for the *H*, *C*, *E*, *D*, and *S* parameters. The column *A* and *B* parameters are less certain. The *B* coefficient for the Agilent SB-C8 column was not significant (i.e., the parameter is not significantly different from zero), as well as the A coefficient for the Agilent 300SB-C3, the Varian/Agilent C18-A. the Agilent Eclipse Plus C18 and the Agilent SB-C8 columns. After omitting these columns from the percent relative error calculations, the percent relative error for the A parameter was 5.9% and the relative error for the B parameter was 12.3%. This latter relative error is consistent with the previous observation that the effect of solute hydrogen bond acidity on retention is mixed (*vida supra*). Overall, less significance should be given to the aB term in Eq. (5). In general, these results suggest that the column parameters can be reported to two digits past the decimal point, and the final column parameter scales are provided in the Supplemental Information as an Excel spreadsheet, along with the standard error of the column parameters calculated as described above.

To evaluate the precision of the solute parameters, normally distributed random errors were added to the column parameters, using the standard error as described above as the scaling factor over 500,000 repetitions. For each repetition, the solute parameters were calculated, and the means and standard deviations of the parameters over the repetitions were determined. The average percent relative standard deviations for each parameter relative to the range of the parameters are shown in Table 4. The e parameter has the largest average error at 5.8%. In general, these results suggest that the solute parameters can be reported to three digits past the decimal point, and the final parameter scales and the corresponding standard errors are provided in the Supplemental Information as an Excel spreadsheet (“Final Parameters for HSM3.xlsx”).

Table 4. Average % relative standard deviations of the solute parameters relative to the range of each parameter[a]

Parameter	**% RSD**
h	1.5
k	2.9
e	5.8
d	4.0
a	2.3
b	3.4
s	2.7

[a] Calculated from 500,000 Monte Carlo iterations as described in the text.

3.3 Isomer selectivity

One goal of the present work was to examine how well the model of Eq. (5) (or equivalently, the 7 PC model), was able to predict the chromatographic selectivity of positional and geometric isomers. Figure 7 shows the predicted selectivity for four positional isomer pairs. The retention order is always predicted correctly. The standard errors for the selectivity predictions over the 13 columns for each isomer pair are shown in Table 5, along with predictions based on HSM1 and HSM2, as described above. Interestingly, selectivities for the cresol isomers and the naphthol isomers are predicted quite well for all three models, with a standard error in α on the order of 0.01. None of these compounds are in the original training set for HSM1 and HSM2. These are relatively simple compounds, with the cresols having a methyl and hydroxyl substitution on the benzene ring, and the naphthols with a hydroxyl substitution on naphthalene. In contrast, the dinitrophenols have two nitro and one hydroxyl groups, and the dihydroxy naphthalenes have two hydroxyl groups. In this case, selectivities predicted by the HSM3 model are improved relative to the HSM1 and HSM2 models (see Table 5).

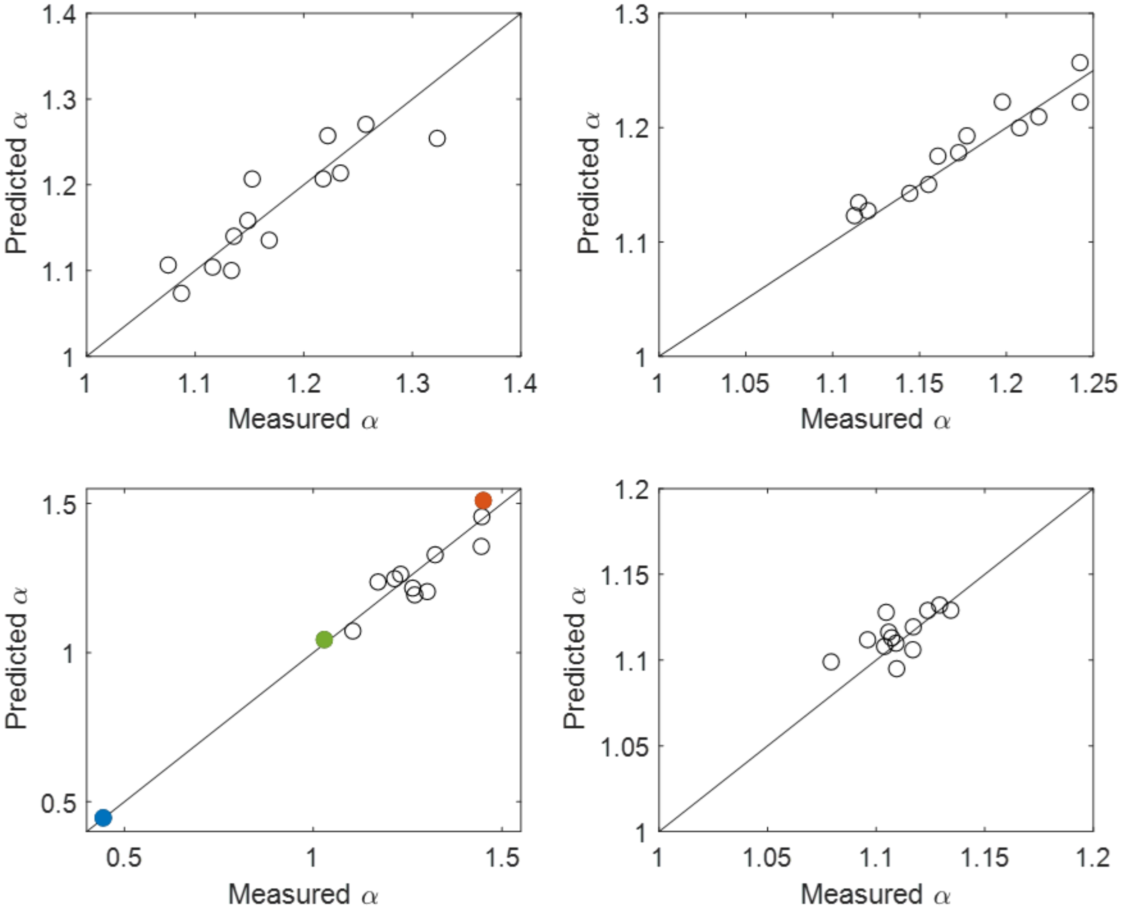


Figure 7. Predicted α value vs. actual α value for (A) 1,2-dihydroxynaphthalene relative to 1,3-dihydroxynaphthalene; (B) 1-naphthol relative to 2-naphthol; (C) 2,5-dinitrophenol relative to 2,4-dinitrophenol; (D) *o*-cresol relative to *p*-cresol. The blue point in (C) corresponds to the selectivity on Bonus RP, the green point corresponds to the selectivity on CSH Phenyl-Hexyl and the orange point corresponds to the selectivity on SB-C18.

Table 5. Positional isomer selectivity standard errors

Selectivity	**HSM1**	**HSM2**	**HSM3**
$a_{1,2\text{-DHN}/1,3\text{-DHN}}$[a]	0.0440	0.0568	0.0319
$a_{1\text{-naphthol}/2\text{-naphthol}}$	0.0135	0.0145	0.0137
$a_{2,5\text{-DNP}/2,4\text{-DNP}}$[b]	0.103	0.175	0.0530
$a_{o\text{-cresol}/p\text{-cresol}}$	0.0110	0.00942	0.0118
Overall	0.0565	0.0922	0.0322

[a]DHN – dihydroxynaphthalene; [b]DNP – dinitrophenol

Of these isomer pairs, the 2,5-dinitrophenol/2,4-dinitrophenol (2,5-DNP/2,4-DNP) pair has an interesting selectivity pattern, as seen in Fig. 7C. The 2,5-DNP is always retained longer than the 2,4-DNP, except on the Bonus RP column. To examine this effect more closely, we compared the contribution of each of the linear terms in Eq. (5) to the calculated ln α for the Bonus RP column ($\alpha_{2,5\text{-DNP}/2,4\text{-DNP}} = 0.45$), the CSH Phenyl-Hexyl column ($\alpha_{2,5\text{-DNP}/2,4\text{-DNP}} = 1.04$) and the SB-C18 column ($\alpha_{2,5\text{-DNP}/2,4\text{-DNP}} = 1.51$). This comparison is shown in Fig. 8. The signs and magnitudes of the *hH*, *kC*, *eE*, *dD*, *bA* and *sS* terms are not significantly different for the three columns. In contrast, the *aB* term is very different on these three columns. The biggest difference in the isomer parameters relative to the parameter range is in the *a* hydrogen bonding parameter (*a* = 0.223 for 2,5-dinitrophenol vs. *a* = 0.349 for 2,4-dinitrophenol, a 29 % difference). This can be rationalized by noting that the nitro groups para and meta to the phenolic oxygen increase the hydrogen bond donating ability of the phenolic group in 2,4-dinitrophenol. To the best of our knowledge, the Bonus RP stationary phase is the only phase of the 13 phases that contains an embedded amide that can serve as a hydrogen bond acceptor (see Table S4). It also has the highest *B* parameter of all the columns studied (*B* = 3.15). Thus, the present model and parameters can help us to rationalize selectivity of this positional isomer pair.

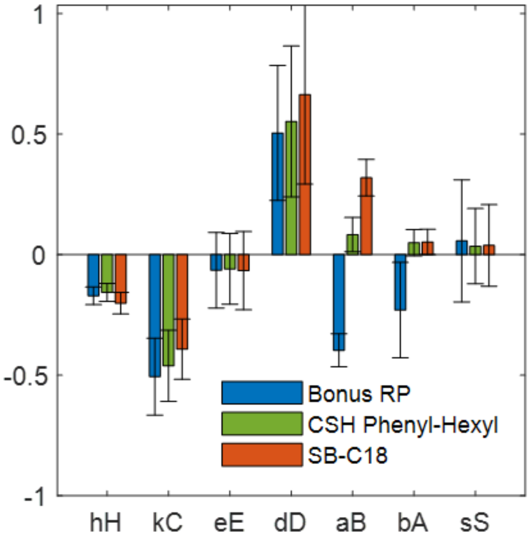


Figure 8. Sign and magnitude of the terms in eq. (5) contributing to the selectivity of 2,5-DNP relative to 2,4-DNP for Bonus RP ($\alpha_{2,5\text{-DNP}/2,4\text{-DNP}} = 0.45 \pm 0.11$), CSH Phenyl-Hexyl ($\alpha_{2,5\text{-DNP}/2,4\text{-DNP}} = 1.04 \pm 0.18$) and SB-C18 ($\alpha_{2,5\text{-DNP}/2,4\text{-DNP}} = 1.51 \pm 0.33$).

The selectivities for the cis- and trans- isomers of stilbene and chalcone are shown in Fig. 9A and 9B. For the stilbenes, we have highlighted the two columns with the largest differences in selectivity, BEH-C8 (blue point in Fig. 9A) and HSS-PFP (orange point in Fig. 9A). The comparison of the contributions of linear terms of Eq. (5) to the selectivity for each column is shown in Fig. 9C. The largest contributor to the difference in selectivity on these two columns is the *eE* term. This can be understood because the *e* parameter for cis-stilbene ($e = 0.080 \pm 0.011$) is less than that of trans-stilbene ($e = 0.134 \pm 0.013$), which is due to the distortion of the double bond in the cis- structure (see 3D structure in Fig. 9E). The polarizability parameter (*E*) for the BEH C8 column is 1.274 ± 0.059, while the *E* parameter for the HSS PFP column is 3.136 ± 0.092.

For the chalcones, the two columns with the biggest difference in trans/cis selectivity are again the BEH-C8 (blue point in Fig. 9B) and HSS-PFP (orange point in Fig. 9B). The corresponding comparison of the linear terms of Eq. (5) is shown in Fig. 9D. In contrast to the stilbenes, several terms make contributions to the selectivity differences for the trans/cis isomers on the BEH-C8 and HSS-PFP columns. In this case, the *hH*, *kC*, *dD*, and *sS* terms all seem to be important contributors. The enhanced trans/cis selectivity on the HSS-PFP column relative to the BEH-C8 is driven by the *hH*, *kC* and *sS* terms, and the *dD* term cancels out some of this selectivity. In this case, it is not as easy to rationalize the difference in selectivities. The differences in solute parameters between the cis- and trans- isomers are quite small, and there are not large differences between their 3D structures as shown in Fig. 9F. It is not expected that either isomer will participate in charge-based interactions, and yet the *kC* term for both columns is significantly different from zero.

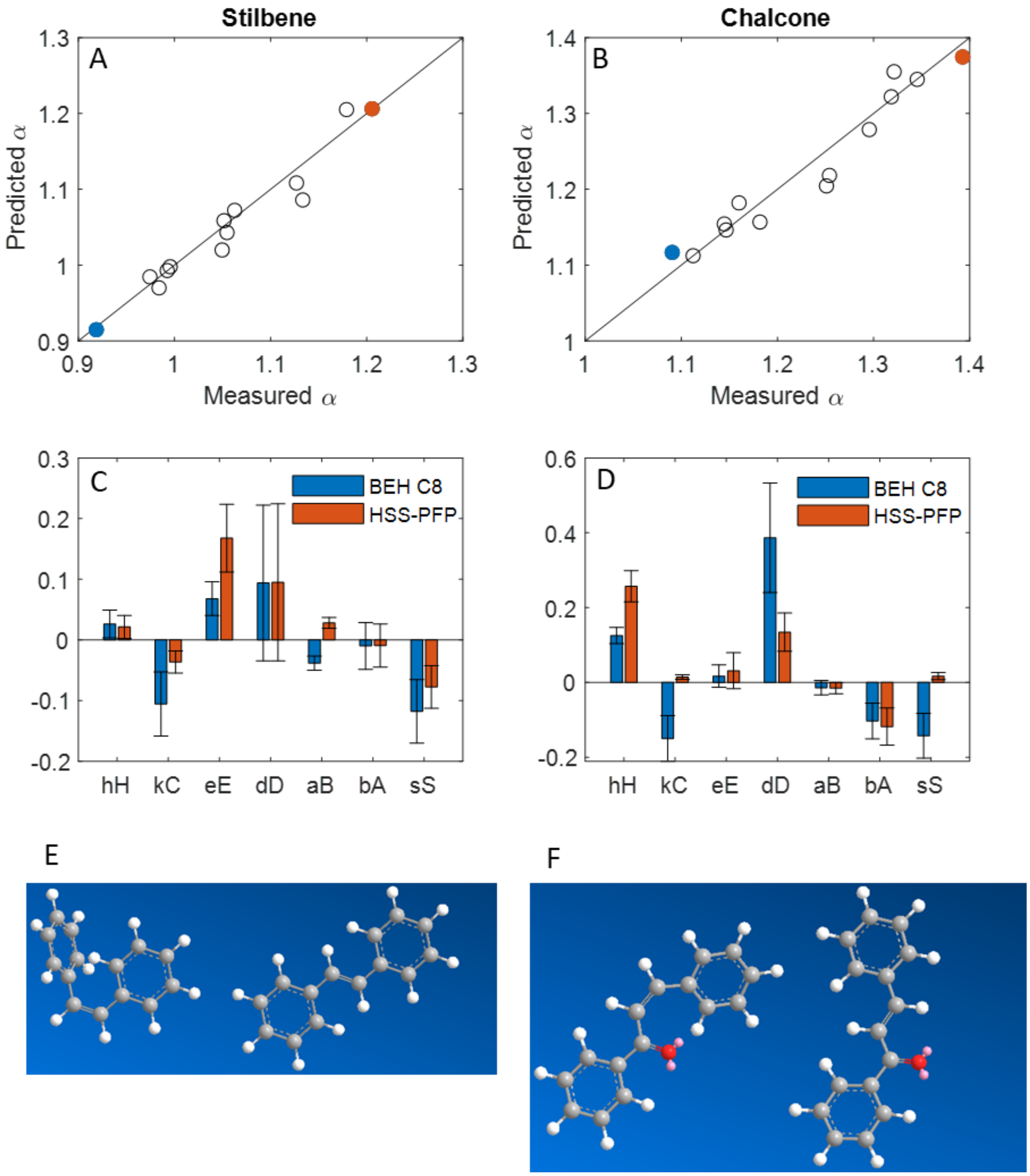


Figure 9. Predicted α value vs. actual α value for (A) trans-stilbene relative to cis-stilbene (blue point is selectivity on BEH C8 and orange point is selectivity on HSS PFP); (B) trans-chalcone relative to cis-chalcone (blue point is selectivity on BEH C8 and orange point is selectivity on HSS PFP); (C) Sign and magnitude of the terms in eq. (5) contributing to the selectivity of trans-stilbene relative to cis-stilbene on BEH C8 ($\alpha_{\text{trans-/cis-stilbene}}$= 0.920 ± 0.023) and HSS PHP ($\alpha_{\text{trans-/cis-stilbene}}$ = 1.202 ± 0.009) (D) Sign and magnitude of the terms in eq. (5) contributing to the selectivity of trans-chalcone relative to cis-chalcone on BEH C8 ($\alpha_{\text{trans-/cis-chalcone}}$ = 1.125 ± 0.037) and HSS PHP ($\alpha_{\text{trans-/cis-chalcone}}$ = 1.378 ± 0.013); (E) Structure comparison of cis-stilbene (left) to trans-stilbene (right); (F) Structure comparison of cis-chalcone (left) to trans-chalcone (right).

The present dataset also includes four pairs of cis/trans isomers related to the Linrodostat pharmaceutical compound (see structures in Fig. 3). The trans/cis selectivities for these compounds are shown in Fig. 10. Note that the retention orders are generally predicted correctly for the Linrodostat related compounds, however the selectivities for the 300SB-C3 (red circles) and SB-C8 (blue squares) columns have larger errors. We did note that for all four isomer pairs, the trans isomer had a larger *s* parameter than the cis isomer, and the 3D models of the cis- isomers showed a more compact structures as compared to the trans- isomers (see Table 6). Two examples of the selectivities observed for these highly similar compounds are shown in Fig. 11. Figure 11A shows the selectivities for Lin-trans-C and Lin-cis-C, and the columns with largest difference in selectivity are highlighted in orange (SB-Phenyl) and blue (Eclipse Plus C18). The linear terms contributing to the selectivity for these two stationary phases are shown in Fig. 11C. For these two isomers, it is not clear what really drives the selectivity because the uncertainties in the individual terms are so large. Figure 11B shows the selectivities for Lin-cis-B and Lin-cis-C, with the selectivities for the Bonus RP and SB Phenyl phases highlighted in blue and orange, respectively. These two compounds share a 6-fluoroquinolin-4-yl)cyclohexyl) core structure (see Fig. 3). The Lin-cis-B molecule contains a tertiary amide, whereas the Lin-cis-C molecule has a secondary amide. This difference is reflected in the *a* parameter, which shows that Lin-cis-C is a stronger hydrogen bond donor (a = 0.104) than cis-B (a = 0.152). In contrast, Lin-cis-B is a stronger hydrogen bond acceptor (b = 0.224) than cis-C (b = 0.124). Because the signs of the *A* and *B* parameters are opposite for the Bonus RP (A = -6.98, B = 2.07) and the SB-Phenyl columns (A = 2.07, B = -1.03), this results in Lin-cis-B being more retained than Lin-cis-C on the SB-Phenyl column, and being less retained than Lin-cis-C on the Bonus RP column, as shown in by the *aB* and *bA* terms in Fig. 11D.

Table 6. Steric parameters (*s*) for Linrodostat and related compounds

	Lin-A	**Lin-B**	**Lin-C**	**Lin-D**
trans isomer	0.451 ± 0.030	0.429 ± 0.028	0.343 ± 0.025	0.248 ± 0.022
cis isomer	0.393 ± 0.027	0.381 ± 0.035	0.258 ± 0.028	0.213 ± 0.030

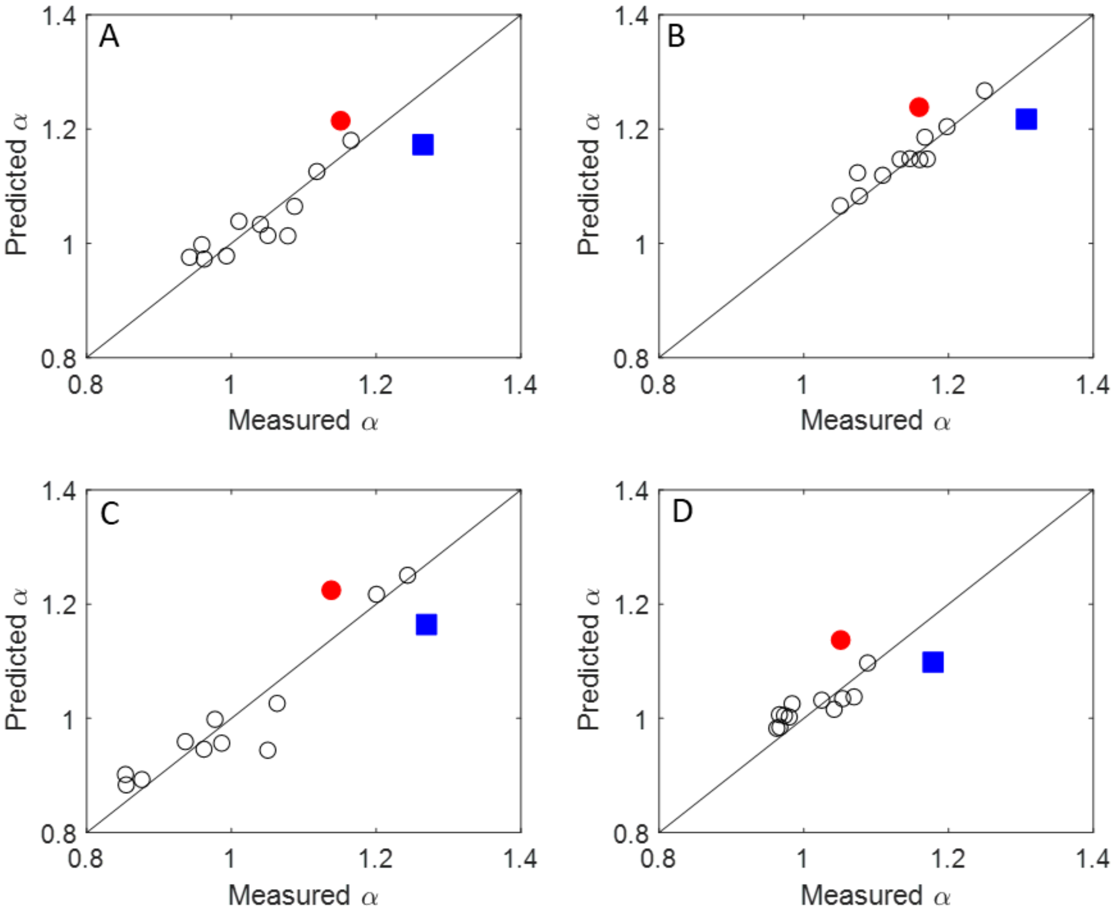


Figure 10. trans/cis isomer selectivities for Linrodostat related compounds. (A) Isomer pair Lin-A; (B) Isomer pair Lin-B; (C) Isomer pair Lin-C; (D) Isomer pair Lin-D. The red circles correspond to the 300SB-C3 column and the blue squares correspond to the SB-C8 column.

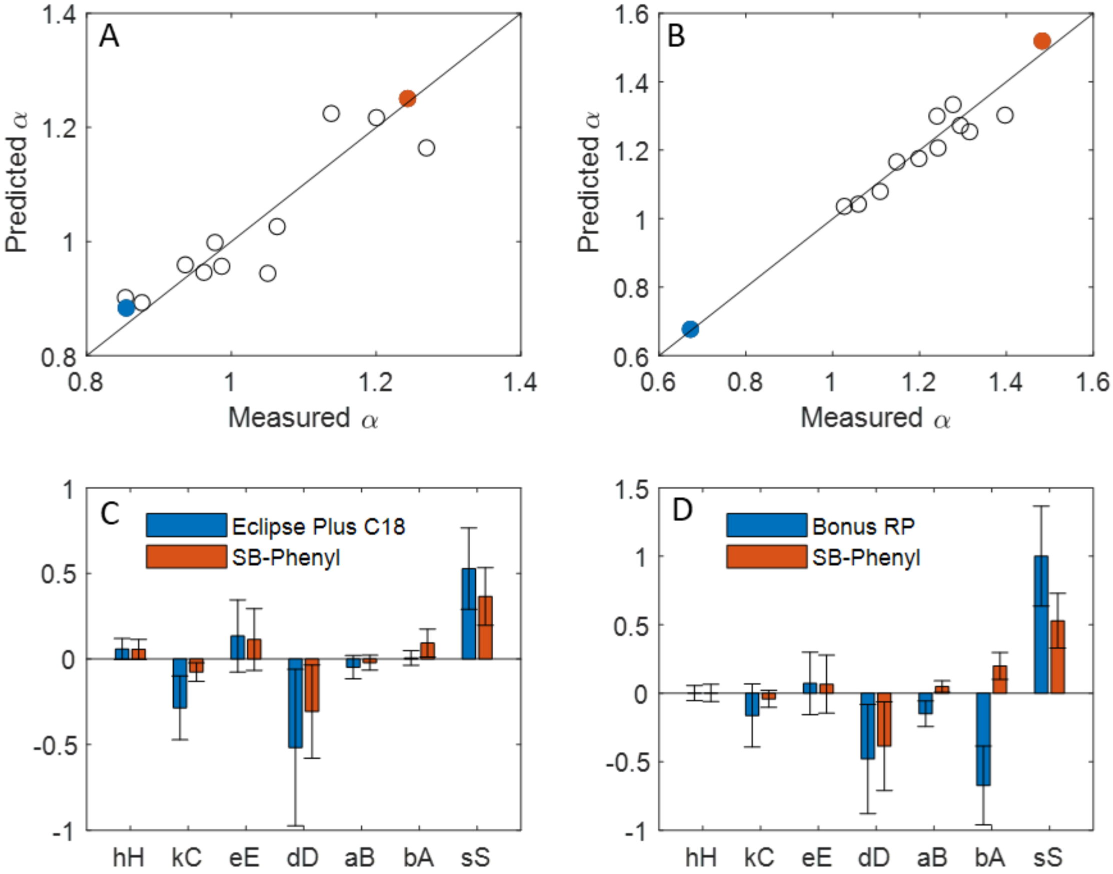


Figure 11. Predicted α value vs. actual α value for (A) Lin-trans-C relative to Lin-cis-C (blue point is selectivity on Eclipse Plus C18 and orange point is selectivity on SB-Phenyl); (B) Lin-cis-B relative to Lin-cis-C (blue point is selectivity on Bonus RP and orange point is selectivity on SB-Phenyl); (C) Sign and magnitude of the terms in eq. (5) contributing to the selectivity of Lin-trans-C relative to Lin-cis-C on Eclipse Plus C18 ($\alpha_{\text{Lin-trans-C-/Lin-cis-C}}$ = 0.88 ± 0.31) and SB-Phenyl ($\alpha_{\text{Lin-trans-C/Lin-cis-C}}$ = 1.25 ± 0.19) (D) Sign and magnitude of the terms in eq. (5) contributing to the selectivity of Lin-cis-B relative to Lin-cis-C on Bonus RP ($\alpha_{\text{Lin-cis-B/Lin-cis-C}}$ = 0.68 ± 0.33) and SB-Phenyl ($\alpha_{\text{Lin-cis-B/Lin-cis-C}}$ = 1.52 ± 0.31.

4. Conclusions

Our recent development of a high throughput approach for acquisition of retention data for liquid chromatography has enabled the collection a large dataset of retention measurements for a varied set of small molecules, many of pharmaceutical significance. The dataset studied here is comprised of 43,329 total measurements made across 13 stationary phases, 89 compounds, and multiple mobile phase compositions. Using a subset of these data, we developed a data-driven model of reversed-phase selectivity based on isocratic retention factors (40% ACN). We refer to the resulting model as HSM3 because it has qualitative characteristics that are similar to the original Hydrophobic Subtraction Model developed by Snyder and coworkers [14].

Our major conclusions drawn from this work follow:

1) Using root-mean-square error of cross validation (RMSECV) to guide development of the model, we found that seven terms were warranted without overfitting the data. Each is a simple linear term composed of a solute property parameter, and a corresponding stationary phase parameter (e.g,. hydrogen bond acidity of the solute paired with hydrogen bond basicity of the stationary phase). Although the parameters originate from a principal components analysis, we have rotated the PCA axes so that they correlate with physicochemical properties that are believed to influence selectivity in RPLC, such as solute hydrophobicity, charge state, and dipolarity.

2) The retention dataset was divided into training and validation subsets. The standard errors in ln α for the fits of the model to these subsets were about 0.033, which roughly corresponds to an average residual from the fit of about 3% in α.

3) The predictive accuracy of HSM3 for the selectivities for a number of isomer pairs appears to be much better than previous models (HSM1 and HSM2).

4) Perhaps most interestingly, an examination of the quantitative contributions of each of the terms in the HSM3 model to the selectivity showed that in some cases the major driver of a separation of closely-related compounds can be identified (e.g., hydrogen bonding). This is a very exciting result in that it may provide a means to de-risk method development by focusing on stationary phase properties that are critical to method robustness, and monitoring those over time.

In our view this work highlights the point that a more detailed understanding of selectivity in liquid chromatography can be realized if we have access to large datasets that span multiple stationary phase and solute chemistries. The ability to use the HSM3 model to rationalize the physicochemical drivers for the separation of specific closely-related solute pairs is very promising, however this work also shows that there most definitely are currently limits to this kind of analysis, as indicated by large uncertainties in some of the terms in the model (i.e., Eq. 5) for specific solute/stationary phase pairs. Much more work is needed to understand the drivers of this uncertainty (e.g., stationary phase drift over time [26]) so that we can work to minimize it in the future.

Acknowledgements

All of the Agilent LC columns and instrumentation used in this work were generously provided by Agilent Technologies. We especially thank Dr. Konstantin Shoykhet for technical assistance with the implementation of feed injection used to collect the experimental retention data. DS, TK, and ZK were supported by funding from the National Science Foundation (CHE2003734).

References

[1] R.I.J. Amos, P.R. Haddad, R. Szucs, J.W. Dolan, C.A. Pohl, Molecular modeling and prediction accuracy in Quantitative Structure-Retention Relationship calculations for chromatography, TrAC - Trends in Analytical Chemistry 105 (2018) 352–359. https://doi.org/10.1016/j.trac.2018.05.019.

[2] P. Kumari, T. Van Laethem, P. Hubert, M. Fillet, P.Y. Sacré, C. Hubert, Quantitative Structure Retention-Relationship Modeling: Towards an Innovative General-Purpose Strategy, Molecules 28 (2023). https://doi.org/10.3390/molecules28041696.

[3] G. Sagandykova, B. Buszewski, Perspectives and recent advances in quantitative structure-retention relationships for high performance liquid chromatography. How far are we?, TrAC - Trends in Analytical Chemistry 141 (2021). https://doi.org/10.1016/j.trac.2021.116294.

[4] S. Baskaran, Y.D. Lei, F. Wania, A database of experimentally derived and estimated octanol-air partition ratios (K0A), J. Phys. Chem. Ref. Data 50 (2021) 043101. https://doi.org/10.1063/5.0059652.

[5] M. Vitha, P.W. Carr, The Chemical Interpretation and Practice of Linear Solvation Energy Relationships in Chromatography, J. Chromatogr. A 1126 (2006) 104–143. https://doi.org/10.1016/j.chroma.2006.06.074.

[6] A. Wang, P.W. Carr, Comparative Study of the Linear Solvation Energy Relationship, Linear Solvent Strength Theory, and Typical Conditions Model for Retention Prediction in Reversed-Phase Liquid Chromatography, J.Chromatogr.A 965 (2001) 3–23.

[7] A.S. Wang, L.C. Tan, P.W. Carr, Global Linear Solvation Energy Relationships for Retention Prediction in Reversed-phase Liquid Chromatography, J.Chromatogr.A 848 (1999) 21–37.

[8] L. Choo Tan, P.W. Carr, M.H. Abraham, Study of retention in reversed-phase liquid chromatography using linear solvation energy relationships I. The stationary phase, J Chromatogr A 752 (1996) 1–18.

[9] E. Tyteca, M. Talebi, R. Amos, S.H. Park, M. Taraji, Y. Wen, R. Szucs, C.A. Pohl, J.W. Dolan, P.R. Haddad, Towards a chromatographic similarity index to establish localized quantitative structure-retention models for retention prediction: Use of retention factor ratio, J Chromatogr A 1486 (2017) 50–58. https://doi.org/10.1016/j.chroma.2016.09.062.

[10] N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, P.W. Carr, Column Selectivity in Reversed-phase Liquid Chromatography. I. A General Quantitative Relationship, J. Chromatogr. A 961 (2002) 171–193. https://doi.org/10.1016/S0021-9673(02)00659-3.

[11] N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, P.W. Carr, Column selectivity in reversed-phase liquid chromatography: II. Effect of a change in conditions, J Chromatogr A 961 (2002) 195–215. https://doi.org/10.1016/S0021-9673(02)00660-X.

[12] N.S. Wilson, J.W. Dolan, L.R. Snyder, P.W. Carr, L.C. Sander, Column selectivity in reversed-phase liquid chromatography: III. The physico-chemical basis of selectivity, J Chromatogr A 961 (2002) 217–236. https://doi.org/10.1016/S0021-9673(02)00658-1.

[13] J.J. Gilroy, J.W. Dolan, L.R. Snyder, Column selectivity in reversed-phase liquid chromatography IV. Type-B Alkyl Silica Columns, J. Chromatogr. A 1000 (2003) 757–778.

[14] L.R. Snyder, J.W. Dolan, P.W. Carr, The hydrophobic-subtraction model of reversed-phase column selectivity, J Chromatogr A 1060 (2004) 77–116. https://doi.org/10.1016/j.chroma.2004.08.121.

[15] D.H. Marchand, L.R. Snyder, J.W. Dolan, Characterization and Applications of Reversed-Phase Column Selectivity Based on the Hydrophobic-Subtraction Model, J Chromatogr A 1191 (2008) 2–20. https://doi.org/10.1016/j.chroma.2007.10.079.

[16] Y. Zhang, P.W. Carr, A visual approach to stationary phase selectivity classification based on the Snyder-Dolan Hydrophobic-Subtraction Model., J Chromatogr A 1216 (2009) 6685–94. https://doi.org/10.1016/j.chroma.2009.06.048.

[17] S. Dragovic, E. Haghedooren, T. Németh, I.M. Palabiyik, J. Hoogmartens, E. Adams, Evaluation of two approaches to characterise liquid chromatographic columns using

pharmaceutical separations., J Chromatogr A 1216 (2009) 3210–6. https://doi.org/10.1016/j.chroma.2009.02.023.

[18] L.R. Snyder, J.W. Dolan, D.H. Marchand, P.W. Carr, The Hydrophobic-Subtraction Model of Reversed-Phase Column Selectivity, in: Advances in Chromatography, Vol. 50, CRC Press, Boca Raton, FL, 2012: pp. 297–376.

[19] A.R. Johnson, C.M. Johnson, D.R. Stoll, M.F. Vitha, Identifying orthogonal and similar reversed phase liquid chromatography stationary phases using the system selectivity cube and the hydrophobic subtraction model., J Chromatogr A 1249 (2012) 62–82. https://doi.org/10.1016/j.chroma.2012.05.049.

[20] J.W. Dolan, L.R. Snyder, The hydrophobic-subtraction model for reversed-phase liquid chromatography: A reprise, LCGC North America 34 (2016) 730–741.

[21] PQRI Database, (2020). https://apps.usp.org/app/USPNF/columnsDB.html.

[22] Column Selectivity Database, (2020). http://www.hplccolumns.org/database/index.php.

[23] T. Dahlseid, A. Florea, G. Schulte, K. Cash, X. Xu, P. Tattersall, Q. Wang, D. Stoll, Changes in the cis-trans isomer selectivity of a reversed-phase liquid chromatography column during use with acidic mobile phase conditions, J Chromatogr A 1708 (2023) 464371. https://doi.org/10.1016/j.chroma.2023.464371.

[24] D.R. Stoll, T.A. Dahlseid, S.C. Rutan, T. Taylor, J.M. Serret, Improvements in the predictive accuracy of the hydrophobic subtraction model of reversed-phase selectivity, J Chromatogr A 1636 (2020) 461682. https://doi.org/10.1016/j.chroma.2020.461682.

[25] D.R. Stoll, G. Kainz, T.A. Dahlseid, T.J. Kempen, T. Brau, B.W.J. Pirok, An approach to high throughput measurement of accurate retention data in liquid chromatography, J. Chromatogr. A 1678 (2022) 463350. https://doi.org/10.1016/j.chroma.2022.463350.

[26] T. Kempen, T. Dahlseid, T. Lauer, A. Florea, I. Aase, N. Cole-Dai, S. Kaur, C. Southworth, K. Grube, J. Bhandari, M. Sylvester, R. Schimek, B. Pirok, S. Rutan, K. Shoykhet, D. Stoll, Characterization of a high throughput approach for large scale retention measurement in liquid chromatography, J. Chromatogr. A 1705 (2023) 464182. https://doi.org/10.1016/j.chroma.2023.464182.

[27] Kempen T, Dahlseid T, Lauer T, Florea A, Aase I, Cole-Dai N, S. Kaur, C. Southworth, K. Grube, J. Bhandari, M. Sylvester, R. Schimek, B. Pirok, S. Rutan, D. Stoll, Characterization of a high throughput approach for large scale retention measurement in liquid chromatography, ChemRxiv (2023).

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

[29] P. Žuvela, M. Skoczylas, J.J. Liu, T. Baczek, R. Kaliszan, M.W. Wong, B. Buszewski, K. Héberger, Erratum: Column characterization and selection systems in reversed-phase high-performance liquid chromatography (Chemical Reviews (2019) 119:6 (3674-3729) DOI: 10.1021/acs.chemrev.8b00246), Chem Rev 119 (2019) 4818. https://doi.org/10.1021/acs.chemrev.9b00167.

[30] N. Ulrich, S. Endo, T.N. Brown, N. Watanabe, G. Bronner, M.H. Abraham, K.-U. Goss, UFZ-LSER database v 3.2.1, (2017). http://www.ufz.de/lserd.

[31] S.C. Rutan, K. Cash, D.R. Stoll, Experimental design and re-parameterization of the Neue-Kuss model for accurate and precise prediction of isocratic retention factors from gradient measurements in reversed phase liquid chromatography, J Chromatogr A 1711 (2023). https://doi.org/10.1016/j.chroma.2023.464443.

[32] Z. Liu, Y. Zhou, Q. Wang, J.P. Foley, D.R. Stoll, J.G. Shackman, Development of tandem-column liquid chromatographic methods for pharmaceutical compounds using simulations based on hydrophobic subtraction model parameters, J Chromatogr A 1695 (2023). https://doi.org/10.1016/j.chroma.2023.463925.

[33] Y. Xu, R. Goodacre, On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning, J Anal Test 2 (2018) 249–262. https://doi.org/10.1007/s41664-018-0068-2.

[34] R. Kiralj, M.M.C. Ferreira, Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application, J. Braz. Chem. Soc 20 (2009) 770–787.

[35] K.B. Sentell, J.G. Dorsey, Retention Mechanisms in Reversed-Phase Liquid Chromatography. Stationary-Phase Bonding Density and Partitioning, Anal Chem 61 (1989) 930–934. https://doi.org/10.1021/ac00184a003.

[36] L.C. Sander, S.A. Wise, Shape Selectivity in Reversed-Phase Liquid Chromatography for the Separation of Planar and Nonplanar Solutes, J. Chromatogr. A 656 (1993) 335–351.

[37] L.C. Sander, M. Pursch, S.A. Wise, Shape selectivity for constrained solutes in reversed-phase liquid chromatography, Anal Chem 71 (1999) 4821–4830. https://doi.org/10.1021/ac9908187.

[38] J.L. Rafferty, J.I. Siepmann, M.R. Schure, Influence of bonded-phase coverage in reversed-phase liquid chromatography via molecular simulation I. Effects on chain conformation and interfacial properties., J Chromatogr A 1204 (2008) 11–19. https://doi.org/10.1016/j.chroma.2008.07.037.

[39] U.D. Neue, C.H. Phoebe, K. Tran, Y.F. Cheng, Z. Lu, Dependence of reversed-phase retention of ionizable analytes on pH, concentration of organic solvent and silanol activity, J Chromatogr A 925 (2001) 49–67. https://doi.org/10.1016/S0021-9673(01)01009-3.

[40] D. V McCalley, The challenges of the analysis of basic compounds by high performance liquid chromatography: some possible approaches for improved separations., J Chromatogr A 1217 (2010) 858–80. https://doi.org/10.1016/j.chroma.2009.11.068.

[41] A. Mendez, E. Bosch, M. Roses, U.D. Neue, Comparison of the acidity of residual silanol groups in several liquid chromatography columns, J Chromatogr A 986 (2003) 33–44.