Accuracy of retention model parameters obtained from retention data in liquid chromatography

Tyler Brau[1], Bob Pirok[1,2], Sarah Rutan[3], and Dwight Stoll[1*]

1) Gustavus Adolphus College
2) Van 't Hoff Institute for Molecular Sciences
3) Department of Chemistry, Box 842006

***Corresponding author**: Professor Dwight Stoll; Gustavus Adolphus College; 800 West College Avenue; Saint Peter, MN 56082; dstoll@gustavus.edu

Non-Standard Abbreviations:

DAD – diode array detector

LSS – linear solvent strength

MRPE - mean residual percent error

NK – Neue-Kuss

SSD – sum of squared differences

Keywords: gradient retention factor, Linear Solvent Strength theory, Neue-Kuss, retention model

Abstract

In liquid chromatography (LC), it is often very useful to have an accurate model of the retention factor, *k*, over a wide range of isocratic elution conditions. In principle, the parameters of a retention model can be obtained by fitting either isocratic or gradient retention factor data. However, in spite of many of our own attempts to accurately predict isocratic *k* values using retention models trained with gradient retention data, this has not worked in our hands. In the present study we have used synthetic isocratic and gradient retention data for small molecules under reversed-phase LC conditions. This allows us to discover challenges associated with predicting isocratic *k*'s without the confounding influences of experimental issues that are difficult to model or eliminate. The results indicate that it is not currently possible to consistently predict isocratic retention factors for small molecules with accuracies better than 10%, even when using synthetic gradient retention data. Two distinct challenges in fitting gradient retention data were identified: 1) a lack of 'uniqueness' in the parameters; and 2) an inability to find the global optimum fit in a complex fitting landscape. Working with experimental data where measurement noise is unavoidable will only make the accuracy worse.

1. Introduction

Several aspects of simulation and method development in reversed-phase (RP) liquid chromatography depend on isocratic retention data. For example, most models of RP selectivity that are used to guide column selection (i.e., either to identify columns with similar or dissimilar selectivities) for small molecule separations are built upon selectivity measurements made under isocratic conditions [1–5]. Also, the theories used to make predictions about the effect of injection volume and sample composition on peak shapes (i.e., volume overload and mobile phase / sample solvent mismatch) depend on the availability of isocratic retention data for analytes of interest in solvents corresponding to the sample and mobile phase compositions [6–10]. In contrast, retention models (e.g., DryLab and similar tools) are often built as a part of method development using training data obtained under gradient elution conditions [11–14]. In the course of predicting optimal separation conditions these tools may suggest gradient conditions with very shallow gradient slopes; as these slopes become more and more shallow, they approach isocratic conditions. It is understood that such predictions are error-prone if they involve extrapolation to gradient slopes outside of the scope of the training data [15].

Making isocratic retention measurements directly for the purposes listed above can be time-consuming, because unless the retention behavior of the molecules of interest are already known, many initial experiments will fail due to conditions that produce retention factors that are too low or too high to be useful [16]. It would be incredibly useful in practice to be able to make retention measurements under gradient elution conditions, and from these measurements extract retention model parameters (i.e., fits of the data to retention models such as the Linear Solvent Strength (LSS) model [17], or the Neue-Kuss (NK) model [18]) that can be used to accurately predict retention under both isocratic and gradient conditions. This could potentially save a lot of time in the process of collecting the training data because gradient methods are better suited to mixtures of analytes, and because conditions can be chosen easily that will likely retain analytes that tend to be poorly retained, while also avoiding retention that is too high (i.e., with a gradient running from 5 to 90% ACN). In spite of many attempts to implement this type of scheme over the past decade, this has been largely unsuccessful in our hands. Our experience has been that isocratic predictions made using retention model parameters extracted from retention data collected under gradient elution conditions are always too inaccurate to be useful for practical application (e.g., errors always larger than 1%, and usually much larger than 1%). We have considered a long list of experimental complications that are difficult to capture in retention models that might be compromising our efforts. For example, it is understood that even modern UHPLC pumps do not produce solvent gradients with perfect accuracy and linearity [19,20]. In our experiments, we have chosen conditions that should mitigate these complications (e.g., using 50 mm x 4.6 mm i.d. columns with 5 micron particles to minimize extra-column effects, viscous heating, pressure effects on retention, and gradient non-linearity at low flow rates). However, even under these well controlled conditions, we and others have not been able to regularly obtain accurate isocratic retention predictions (errors < 1%) from training data collected under gradient conditions [21–23]. These experiences have led us to the present work, which aims to understand the factors that lead to inaccurate predictions for isocratic retention factors calculated using retention parameters obtained from fitting gradient retention data. In this work, we use synthetic data so that the study is not affected by experimental complications that are difficult to eliminate, such as gradient delays and distortions.

In this study we first determined isocratic retention factors experimentally for a variety of small molecules under RP conditions using 61 different analyte/stationary phase pairs. Retention parameters were then extracted from these data by fitting them to the NK model. The resulting parameters were treated as "reference retention parameters", and were used to calculate a set of "reference retention factors", k_{ref}. Using the reference retention parameters, different sets of synthetic isocratic or gradient retention data were produced, with or without simulated measurement noise added. Each set of synthetic data was then fit to the NK model to obtain retention parameters using different fitting approaches. In this work, we have focused our attention on the NK model, because in our experience it provides good isocratic predictions for a broad range of molecules and experimental conditions. Finally the retention parameters were used to predict isocratic retention factors, which were then compared to the k_{ref} values and evaluated for their accuracy.

2. Experimental

2.1 Collection of experimental isocratic retention data

2.1.1 – Chemicals

Milli-Q water (18.2 MΩ) was obtained from a Millipore purification system (Burlington, MA). All analyte compounds, ammonium hydroxide (28-30%), formic acid (> 95%), and acetonitrile (ACN) were purchased from Sigma-Aldrich (St. Louis, MO) and used as-is. The cis- isomer of chalcone was obtained by exposing a solution of the trans- isomer in ACN to sunlight at room temperature for one day. Stock solutions for each analyte were prepared at 10 mg/mL in either neat ACN or 50/50 ACN/water. Analytical samples were prepared by diluting the stock solutions to either 0.2 or 5.0 mg/mL using 50/50 ACN/water as needed to give a peak height greater than 10 mAU at 254 nm. The analytes used in this study are listed in the Supplemental Information, Table S1.

2.1.2 – Mobile phase preparation

The aqueous component of the mobile phase, which we refer to as 25 mM ammonium formate, pH 3.2, was prepared gravimetrically in 2-L batches, according to the following recipe. To a 2-L solvent bottle were added 1986.2 g of water, 2.92 g of ammonium hydroxide (29.1%), and 9.92 g of formic acid (97.4%). The solution was used after mixing thoroughly without any further pH adjustment.

2.1.3 – Instrumentation, columns, and conditions

Retention measurements were made using an Agilent HPLC system (Waldbronn, Germany). The system included a binary pump (G4220A) with Jet Weaver V35 Mixer (G4220-68135), autosampler (G7167A), thermostatted column compartment (G1316C), and diode-array detector (DAD) (G4212A) equipped with a Max-Light Cartridge Cell (G4212-60038, 10 mm path length). The system was controlled using Agilent OpenLAB CDS Chemstation Edition (Rev. C.01.10). The injection volume for each analysis was 0.15 µL. Two columns were used in this work: 1) Agilent Zorbax SB-C18 (5 mm x 2.1 mm i.d., 1.8 µm); 2) Agilent Zorbax Bonus RP (5 mm x 2.1 mm i.d., 1.8 µm) (see Table S1).

The flow rate for all measurements was 1.0 mL/min., and the temperature was 40 °C. Mobile phases were “machine-mixed” by the binary pump. A minimum of five mobile phases were used for each compound, where the compositions were chosen such that all retention factors were between 1 and 50, but roughly evenly spaced in that range. Five replicate retention measurements were made at each composition, and the means of these values were used as described in Section 3.1.

2.1.4 – Retention factor calculations

Isocratic retention factors were calculated using Eq. 1, where the column dead time (t_m) and the extra-column time (t_{ex}) were determined using uracil (0.1 mg/mL) in a mobile phase of 50/50 ACN/buffer. We are well aware that these conditions do not provide the most accurate measure of the column dead time [24], however a small inaccuracy in this value will have no effect on the conclusions we draw from the study described in this paper. Based on other work in our laboratory we have made a correction to these *k* values to compensate for the volume of the column frits that is normally unaccounted for in the measurements of t_{ex}, and important when working with columns as small as those used here. This correction amounts to an increase all *k* values of about 20%; details the provide the basis for this correction will be published separately elsewhere, but should have no influence on the conclusions that follow from this study.

$$k = \frac{t_r - t_m}{t_m - t_{ex}} \tag{1}$$

3. Calculations

3.1 Initial fitting of experimental isocratic data

The experimental k vs. ϕ data were fit to the NK model [18], described by the equation

$$\ln(k) = \ln(k_w) + 2\ln(1 + S_2\phi) - \frac{S_1\phi}{1 + S_2\phi} \tag{2}$$

where k_w is the retention factor in pure weak solvent, S_1 is analogous to the slope of ln(*k*) vs. ϕ in LSS theory [17], and S_2 accounts for any curvature in the ln(*k*) vs. ϕ plot. Fitting was performed with the *lsqnonlin* function in MATLAB using the trust-region reflective algorithm, which required an initial guess for each parameter. This initial guess for *lsqnonlin* was generated by setting $S_2 = 0$ and computing the closest straight-line approximation of $\ln(k)$ vs. ϕ to give approximate values for S_1 and k_w. These approximate values were then used with S_2 = 0 as the initial guess for *lsqnonlin*, with the algorithm set to run for a maximum of 1×10^6 iterations, 1×10^6 function evaluations, and to minimize to a function tolerance of 1×10^{-10}. If this procedure did not result in a reasonable fit to the data as measured by correlation coefficients ($R^2 > 0.999$), the initial guess was manually tuned and *lsqnonlin* was run again until a reasonable fit was obtained. In total, 61 sets of parameters were obtained and used as described below. The three-dimensional space occupied by the parameters is shown in Fig. S1. In Fig. 1 (box 2) these parameters are referred to as the "reference" values of S_1, S_2, and k_w. We note that all fitting results reported throughout this paper refer to fits of synthetic data generated using these "reference" values.

3.2 Generation and fitting of synthetic retention data

3.2.1 Isocratic elution

To generate synthetic isocratic retention data for fitting, the reference values of k_w, S_1, and S_2 were used in combination with Eq. 2 to calculate the ϕ values that correspond to k values of 1 (ϕ_{upper}) and 50 (ϕ_{lower}). Ten evenly spaced data points (with respect to ϕ) were then selected between ϕ_{upper} and ϕ_{lower}. Retention factors were calculated using these ϕ values; these are referred to hereafter as k_{ref} (Fig. 1 (box 3)). In cases where noisy synthetic data were used, five replicates at each ϕ value were generated with normally distributed noise with a specified standard deviation, σ. The synthetic $\ln(k)$ vs. ϕ data were then fit to a straight line to give approximate values of $\ln(k_w)$ and S_1 (Fig. 1 (box 5)). These approximate values were then used with S_2 = 0 to provide an initial guess for *lsqnonlin*, which then fit the $\ln(k)$ vs. ϕ data to Eq. 2 (Fig. 1 (box 6)). Errors were calculated as mean residual percent errors (MRPE) using Eq. 3 (Fig. 1 (box 7))

$$MRPE = \frac{\sum_{i=1}^{n}\frac{k_{fit}-k_{ref}}{k_{ref}}\times100}{n} \qquad (3)$$

where n is the number of datapoints, k_{fit} is the retention factor predicted by the fit model (Fig. 1 (box 6)), and k_{ref} is the 'reference' retention factor predicted by the 'reference' parameters obtained from the initial fitting of the experimental isocratic data (Fig. 1 (boxes 2,3)). For each fit, k_{fit} and k_{ref} were calculated for ten points in the range of $\phi_{lower} < \phi < \phi_{upper}$.

3.2.2 Gradient elution

In Fig. 1, the step in box 2 was followed by determination of gradient times (t_g) that would give effective retention factors (k_{eff}) between 1 and 50, where k_{eff} was calculated using Eqs. 4a and 4b [23,25]

$$k_{eff} = \frac{t_D}{t_0} + \frac{\frac{\phi_i + \frac{(1+S_2\phi_i)}{S_1} ln\left[1+\beta k_w S_1\left(t_0 - \frac{t_D}{k_w(1+S_2\phi_i)^2 e^{-\frac{S_1\phi_i}{1+S_2\phi_i}}}\right)e^{-\frac{S_1\phi_i}{1+S_2\phi_i}}\right]}{1-\frac{S_2(1+S_2\phi_i)}{S_1} ln\left[1+\beta k_w S_1\left(t_0 - \frac{t_D}{k_w(1+S_2\phi_i)^2 e^{-\frac{S_1\phi_i}{1+S_2\phi_i}}}\right)e^{-\frac{S_1\phi_i}{1+S_2\phi_i}}\right]} - \phi_i}{\beta t_0} \tag{4a}$$

$$\beta = \frac{\phi_f - \phi_i}{t_g} \tag{4b}$$

where t_D is the gradient delay time, t_0 is the column dead time (0.1 min), ϕ_i is the solvent composition used at the starting point in the gradient (ϕ_i = 0.05 was used here), ϕ_f is the solvent composition used at the endpoint in the gradient (ϕ_f = 0.70 was used here), and t_g is the duration of the gradient. Equation 4a was solved for the gradient times corresponding to k_{eff} = 1 ($t_{g,lower}$) and k_{eff} = 50 ($t_{g,upper}$) for each set of parameters. If the lower bound on k_{eff} could not be reached for a given analyte, then $t_{g,lower}$ was set to be equal to 0.1 min greater than $t_{g,min}$, where $t_{g,min}$ was defined as the shortest gradient time providing elution of the analyte within the gradient time (i.e., $t_r < t_g$), and calculated as

$$t_{g,min} = \frac{(\phi_f - \phi_i)\, k_w\, S_1\left(t_0 - \frac{t_D}{k_w(1+S_2)^2 e^{-\frac{S_1\phi_i}{1+S_2\phi_i}}}\right) e^{-\frac{S_1\phi_i}{1+S_2 * \phi_i}}}{\frac{(\phi_f - \phi_i)S_1}{e^{1+S_2(\phi_i+\phi_f)+S_2{}^2\phi_i\phi_f}-1}} \tag{4c}$$

Similarly, if the upper bound on k_{eff} could not be reached (i.e., $k_{eff,max}$ < 50, given ϕ_i = 0.05), $t_{g,upper}$ was set to be 15 min. Then, ten evenly spaced gradient times were then selected between $t_{g,lower}$ and $t_{g,upper}$, and k_{eff} was calculated using Eq. 4a for each value of t_g. In the case where noise was added to synthetic retention data (Fig. 1 (box 9)), the same procedure was used as described in Section 3.2.1.

Three different strategies were used to fit the synthetic gradient retention data (Fig. 1 (boxes 10a-c/11a-c)), with each yielding a set of S_1, S_2, and k_w values. A summary of the approaches used for fitting both isocratic and gradient data, and the corresponding results, is given in Table 1.

3.2.2.1 - Basic Fitting Procedure (Fig. 1 (box 10a/11a))

The k_{eff} vs. t_g data were fit to Eq. 4a using the *lsqnonlin* algorithm with small positive, non-zero initial guesses for each parameter (i.e., S_1, S_2, k_w all equal to 1 (Fig. 1 (box 10a)).

3.2.2.2 - Global Search Fitting Procedure (Fig. 1 (box 10b/11b))

The k_{eff} vs. t_g data were fit to Eq. 4a using the *GlobalSearch* algorithm in MATLAB, rather than *lsqnonlin*, with bounds for each parameter set at $k_w = 1.0$ to 1.0×10^9, $S_1 = 5.0$ to 400, and $S_2 = 0.05$ to 15. *GlobalSearch* generates a large number of initial guesses within these bounds and evaluates them using the *fmincon* fitting algorithm (a constrained function minimizer using the interior-point approach and 100,000 start points) before returning the best set of fit parameters. This best set was then further refined using the *lsqnonlin* algorithm.

3.2.2.3 – Fitting Procedure with Parameter Scanning (Fig. 1 (box 10c/11c))

The k_{eff} vs. t_g data were fit to Eq. 4a using *lsqnonlin* with S_2 fixed at 0 in order to provide approximate values for k_w and S_1. Multiple fits were then performed using *lsqnonlin* along with the estimates of k_w and S_1, along with multiple values of S_2 in the range of 0-15 at 0.01 unit increments, with the best fit parameters reported at the end.

Following each fitting procedure, errors were evaluated in the same way as described in Section 3.2.1. The S_1, S_2, and k_w values obtained from a fit of gradient retention data were used to calculate isocratic retention factors (k_{fit}; Fig. 1 (box 12)) and compared to the k_{ref} values calculated from the 'reference' parameters (Fig. 1 (box 2)).

4. Results and Discussion

4.1 Fitting noise-free isocratic data

In order to ensure that *lsqnonlin* was an appropriate choice of algorithm for determining parameters for the NK model, the procedure for isocratic fitting described in Fig. 1 (boxes 5, 6) was applied to synthetic isocratic data generated (Fig. 1 (box 3); no noise added) using the 61 parameter sets (Fig. 1 (box 2)). In the Supporting Information we provide several figures that illustrate the characteristics of these fits. Figure S2A shows the fit for 2,2'-dipyridyl, and Fig. S2B shows the fit for benzonitrile. Figures S2C and S2D show the k_{fit} values and the error in the k_{fit} values for all 61 parameter sets. In one case (berberine, SB-C18, see Fig. S2D), errors of about 0.1% in k were observed due to slow progress toward the correct parameters near the minimum of the objective function, rather than convergence to an incorrect set of parameters. Otherwise, given a sufficiently good initial guess (based on a linear approximation of ln(k) vs. ϕ; see Section 3.2.1) and noise-free data, the *lsqnonlin* algorithm consistently converged on correct NK parameters, producing isocratic k_{fit} values (Fig. 1 (box 7)) that were within 0.01% of the k_{ref} values (Fig. 1 (box 3)). In other words, the algorithm works well in the case of fitting noise-free isocratic data. The numerical values of the parameters from each fit are provided as Supporting Information in Table S1.

4.2 Fitting noisy isocratic retention data

Figure S3 shows the impact of adding noise to synthetic isocratic retention data (Fig. 1 (box 4)) on the errors (Fig. 1 (box 7)) obtained from fitting using the basic approach illustrated described in Section 3.2.1 (Fig. 1 (box 6)). As discussed in Section 4.1, Fig. S3A shows that in the absence of noise, an initial guess calculated by simply fitting a straight line to a plot of ln(k) vs. ϕ (Fig. 1 (box 5)) is sufficient for the *lsqnonlin* algorithm to find parameters that enable accurate predictions of isocratic k values with average errors below 0.001%. This is the same type of result shown in Fig. S2, but now for all 61 parameter sets, and the errors plotted in a histogram. This was not a surprising result, as the lack of noise combined with the good initial guess meant that the fitting algorithm was able to reach a high level of accuracy in the fit parameters given enough iterations. This result also provides a baseline against which we can compare results obtained after fitting retention with noise added to synthetic data. Figure S3B shows the effect of adding noise at 0.05%

(see Section 3.2.1; 100*σ/k = 0.05) to the synthetic retention data. While all but one fit resulted in less than 1% average prediction error, the distribution of errors shifted towards larger errors, indicating that adding noise imposes a limit on the accuracy of the predictions. Figure S3C shows that the distribution of errors shifts even further to the right when the added noise is increased to 0.5%. The fraction of fits with >1% average error in predicting isocratic k values rose to 26.2%, indicating that even isocratic fits with a good initial guess are not immune to the effects of increasing levels of measurement noise.

4.3 Fitting noise-free gradient data

Using the same underlying parameters (Fig. 1 (box 2)), a similar investigation was conducted using synthetic gradient retention data instead of synthetic isocratic retention data. Figure 2 shows representative fits for two specific compounds (A - 2,2'-dipyridyl, and B - benzonitrile), and a summary of errors obtained for all 61 fits (Panels C and D). A significant difference between fitting the isocratic (Section 4.1) and gradient (Section 4.3) data is related to the initial guess used to initiate the *lsqnonlin* algorithm. Whereas in the isocratic case (Section 4.1) we were able to compute an initial guess for each fit by first approximating Eq. 2 with a linear relationship, this is not the case when fitting to Eq. 4, as there is no obvious analogous approximation that can be made. Therefore, for each fit leading to the results shown in Fig. 2, the initial guess was chosen to be $k_w = 1.0$, $S_1 = 1.0$, and $S_2 = 1.0$. While Fig. 2B shows that starting with this simple guess led to a fit that resulted in accurate predictions of isocratic k values for some compounds such as benzonitrile (SB-C18), Fig. 2A shows that the same procedure fails significantly in other cases (2,2'-dipyridyl; SB-C18). Figures 2C/D show how extensive the disagreement between k_{fit} and k_{ref} was in general, as the percent error in k ranged anywhere from about 0 to 10,000%. The numerical values of the parameters from each fit are provided as Supporting Information in Table S2. Specifically, for the 61 cases investigated, 27 cases yielded parameters within 0.1 % of the 'reference' parameters; the remaining cases resulted in parameters that showed dramatic deviations from the 'reference' values. This is a very important result. Even though we have started with the same parameter sets (Fig. 1 (box 2)) for fitting the isocratic (Section 4.1) and gradient (Section 4.3) data, fitting noise-free isocratic retention data consistently yields highly accurate retention parameters, whereas fitting noise-free gradient retention data does not, at least when a simple initial guess is used. In the case of the gradient data, the resulting fits are so bad they cannot be trusted at all.

4.3.1 Effect of Initial Guess on Fitting Gradient Retention Data

Figure 3 shows the impact of using different initial guesses on the quality of fitting noise-free synthetic gradient retention data using the *lsqnonlin* algorithm (Fig. 1 (box 11a). Each panel in Fig. 3 is a histogram of the errors (i.e., average difference between k_{ref} (Fig. 1 (box 8)) and k_{fit} (Fig. 1 (box 12)); logarithmic x-axis) obtained for all 61 parameter sets (Fig. 1 (box 2)) using initial guesses that were different from the reference parameters by a multiplier α. Panels A and B show the most and least challenging cases, with α equal to 1 x 10^{-8} or unity, respectively. In other words, Panel A shows the errors obtained when the initial guess is $k_w = 1 \text{ x } 10^{-8} \times k_{w,ref}$, $S_1 = 1 \text{ x } 10^{-8} \times S_{1,ref}$, and $S_2 = 1 \text{ x } 10^{-8} \times S_{2,ref}$, and Panel B shows the errors obtained when the fitting procedure is initiated with the same parameters (Fig. 1 (box 2)) used to produce the synthetic retention data. The results in Panel B show that the fitting procedure works correctly when the initial guess is very close to the correct solution. On the other hand, starting with guesses near zero for all three parameters yields poor results (i.e., errors >> 1%) for 50% of the observations. Figure 3C, shows that increasing the values of the initial guess just 10% beyond the reference values (i.e., α = 1.1) resulted in errors larger than 1% about 9% of the time. Scaling the initial guess even farther from the reference parameters resulted in larger errors, as one might expect. Figure 3D shows results for guesses with α = 2, where we see that 18% of the fits produced average errors larger than 1%. These results make it clear that the initial guess provided to the fitting algorithm plays a critical role in determining whether or not the *lsqnonlin* algorithm converges on the correct NK parameters.

In attempt to understand why the initial guess influenced the accuracy of fitting gradient retention data using the *lsqnonlin* algorithm so strongly, visualizations of the fitting space were constructed. Figure 4 shows three-dimensional plots (one for amitriptyline, and one for 2,2'-dipyridyl) of the value of the sum of squared differences (SSD) between the gradient k_{eff} values calculated (Eq. 4) using the 'reference' fitting parameters (Fig. 1 (box 2)) and the k_{eff} values calculated using the parameters indicated by a point in the space (i.e., each point in the space represents a possible combination of k_w, S_1, S_2 the algorithm may encounter in fitting the data). Each data point is colored to indicate the value of the SSD at that point, with the color applied on a logarithmic scale given the many orders magnitude spanned by the objective function. In both Fig. 4A and Fig. 4B we observe multiple, broad regions populated by relatively low SSD values separated by "sheets" of large SSD values, which we refer to as "barriers" that fitting algorithms must get over or through on the way to finding the global minimum. For Fig. 4A, the maxima in the SSD values populate a curved surface (yellow/orange band), which cut the fitting space into two regions where the correct parameters for amitriptyline were located on one side of the surface and the other side is quite "flat" with no major barriers populated by large SSD values. A band of low SSD values (dark blue) was also present in the plot, which includes the values of the 'reference' parameters (Fig. 1 (box 2)). The maxima acted as a barrier to the fitting algorithm, as any initial guess placed in the region opposite to the one that contained the 'reference' parameters always resulted in a fit that moved away from the correct minimum, as the algorithm will always move in a direction that decreases the value of the sum of squared differences. Likewise, an initial guess placed in a region of uniform color was not likely to converge on the correct set of parameters as the objective function was flat in that space – while it was not flat in a three-dimensional sense, it was flat in a four-dimensional one, as moving to any point nearby in the fitting space did not cause a significant change in the value of the sum of squared differences. For Fig. 4B, the fine structure of the maxima are more obvious due to the scaling of the plot. The maxima formed shells that split the fitting space into multiple regions, with the 'reference' parameters for 2,2'-dipyridyl only located within one of the shells. For this parameter set, any initial guess that was made would have converged on the minimum of the SSD in the corresponding shell and returned the location of this local minimum as the fit parameters. In this case, finding the correct, 'true' parameters is highly unlikely, as this would require the algorithm to get over/through multiple barriers, whereas in the case of Fig. 4A there is only one major barrier involved.

Figure 5 shows the mean percent difference between isocratic retention factors k_{ref} and k_{fit} obtained by fitting noise-free synthetic gradient retention data using *lsqnonlin* with the indicated point in the three-dimensional space as the initial guess. For Fig. 5A, three distinct regions are observed: 1) initial guesses that result in fits with negligible error in k_{fit} (dark blue points); 2) initial guesses that result in a large amount of error (red points); and 3) initial guesses that result in a (relatively) moderate amount of error (light blue points). The boundary separating the regions of high and low errors mirrors the location of the maxima in the SSD plot in Fig. 4A, confirming that the fitting algorithm could not penetrate the barrier in the objective function that separates the parameter space into two main parts. It is also clear that on each side of the barrier there are multiple local minima in the objective function, and that these minima lead to very different levels of prediction error (10^5% and 10^{20}%). However, Fig. 5A also shows that being on the same side of the barrier as the 'true' parameters was not sufficient to guarantee convergence to the 'true' parameters, as starting with an initial guess located too far from the 'true' parameters sometimes yielded prediction errors greater than 10^5%. This was likely due to the SSD surface being flat in this region, as indicated by Fig. 4A, which prevented the fitting algorithm from making significant progress toward the correct parameters. Note that in the presence of noise, the effective 'flatness' of the SSD surfaces will be enhanced, causing even more difficulties in converging to the k_{ref} values. Figure 5B shows the same type of mean isocratic retention factor error plot, but for 2,2-pydridyl. Comparing Figs. 4B and 4B we see a similar mirroring of the characteristics in these plots that we observed with Figs. 4A and 5A. Whereas the boundary between the regions of low and high error in Fig. 8A closely resembled a plane, in Fig. 5B we see a shell-like structures similar to those in Fig. 4A where the magnitude of the prediction error depended on which shell the initial guess was located. The region that produced the lowest prediction error was again the shell that contained the 'reference' parameters, while initial guesses located in any other shell resulted in prediction errors that ranged from 10^5 to 10^{15}%. The largest errors corresponded to the initial guesses located close to the k_w axis (where the initial guess for S_1 approaches 0). Note that the basins of convergence for both Fig. 5A and Fig. 5B (regions where the prediction error is negligible) did not conform to a simple geometric shape, making a useful mathematical description of the shape of these regions difficult. Manual inspections of the parameter landscapes in Figs. 4 and 5 showed that the regions corresponding to very large errors often involve combinations of S_1, S_2, and k_w that lead to chromatographically unrealistic outcomes. In principle the apparent barriers in the fitting landscape could be avoided by preventing the fitting algorithm from evaluating combinations of parameters that lead to chromatographically unrealistic outcomes, but this would eliminate the possibility of unsupervised fitting, and at this point it time we do not know how transferrable the behavior illustrated in Figs. 4 and 5 are to other compounds, columns, and conditions. This is an area of ongoing study. Readers interested in the fine structure of the cubes in Fig. 5 are referred to movies provided as Supplemental Information that have been constructed by viewing one slice of the cube at a time (see Section S3).

Figure 6 gives some insight as to what the barriers in Fig. 4 and the various regions of error in Fig 5 corresponded to in terms of fit quality. Figure 6A shows the final fit (black line) to the synthetic gradient retention data (red points) for 2,2'-dipyridyl, while Fig. 6D shows the corresponding comparison of isocratic predictions (k_{fit}) to k_{ref}. The initial guesses used in these cases were k_w = 1.0, S_1 = 1.0, and S_2 = 1.0, which fell into a region with approximately 10^6% error. Figures 6B/E show the results obtained when the initial guess was shifted to k_w = 7.0, S_1 = 7.0, and S_2 = 0.0. Although this initial guess was only 9% closer to the 'reference' parameters, the average error in isocratic predictions decreased by 15 orders of magnitude to 10^{-9}%.

Given these results, it is clear that other algorithms that are designed to more comprehensively sample the parameter space are worth exploring. Among several algorithms we have tried for this purpose, the Matlab *GlobalSearch* algorithm has performed the best in our hands; the results of this work are described below in Section 4.4. Finally, in Figures 9C/F, we show the case for initial guesses of k_w = 115, S_l = 7.8, and S_2 = 15, Here, the fit to k_{eff} is particularly bad throughout the range. While a fit quality metric would lead to rejecting this result, cases with a large number of experiments could result in many poor fits. This would lead to much lower data analysis throughput and more manual intervention to obtain adequate fits (with results that still may not be unique).

4.4 Fitting noisy gradient retention data

After establishing a baseline for the performance of the basic approach for fitting synthetic retention data as shown in Fig. 6, the performance of the basic approach for fitting synthetic gradient retention data with noise added was assessed using an initial guess of k_w = 1.0, S_1 = 1.0, and S_2 = 1.0; these results are shown in Fig. 7. Figure 7A confirms the results discussed earlier in Section 4.2 for noise-free synthetic gradient retention data; 47.5% of fits resulted in average errors larger than 10% for prediction of isocratic retention. This compares to 100% of fits yielding prediction errors less than 1% when predictions are made based on fits of isocratic data (i.e., Fig. S3A). As was the case in Fig. S3, adding noise to synthetic gradient retention data shifts the distributions of errors to the right (Figs. 7B/C), making a bad situation even worse. Again, the higher the noise level, the further the shift of the error distribution to larger errors. At a relative noise level of 0.5% (Fig. 7C), 95.1% of fits produced average isocratic prediction errors larger than 1%. When compared to the basic approach for fitting isocratic retention data, the basic approach to fitting gradient retention data performs much worse at any level of noise. This is partly due to the fact that a reasonable initial guess can be estimated when fitting isocratic data using Eq. 2, while no such option is available for Eq. 4 due to its complexity. Another challenge is that gradient retention data are oftentimes not as ‘unique’ as isocratic retention data. While an individual gradient retention measurement may span a range of ϕ values compared to a single isocratic measurement, the effective retention factor is fundamentally an integrated quantity dependent on the mobile phase history experienced by the analyte up to the point in time that it exits the column. This can result in a situation where two compounds with very different retention histories (i.e., different mobile phase experiences) can wind up eluting with exactly the same k_{eff} value. As a result, it becomes necessary to thoroughly search the NK parameter space in order to get consistently accurate results.

One approach to address this challenge is to use a different fitting algorithm. Figure 8 shows the performance of the *GlobalSearch* algorithm for fitting synthetic gradient retention data with the same levels of noise as in Fig. S3. For noise-free retention data (Fig. 8A), *GlobalSearch* returned parameter sets that yield isocratic predictions with less than 1% average error, which represents a significant improvement over the basic fitting approach (i.e. compare Fig. 8A to Fig. 7A). However, when noise is added to the synthetic retention data at the level of 0.05%, the percentage of fits yielding average isocratic predictions errors below 1% error falls to just 41%, with 53% of fits producing errors between 1 and 10%. At a relative noise amplitude of 0.5%, only 4.9% of fits produced parameters that yielded isocratic predictions better than 1% on average.

While the *GlobalSearch* algorithm approach to fitting noise-free gradient retention data did not perform as well as the basic approach to fitting noise-free isocratic retention data, it did offer a significant improvement over the basic approach to fitting noise-free gradient retention data, providing parameters that yielded isocratic predictions with better than 1% average error in an additional 43% of cases. However, this improvement was diminished as the level of noise added to the synthetic retention data was increased. Relative to the basic approach to fitting isocratic retention data, the *GlobalSearch* algorithm was much more susceptible to the influence of noise.

We again attribute this to the lack of ‘uniqueness’ in the gradient retention data – while thoroughly searching the fitting space could result in the correct answer in most cases where noise is absent, adding noise at even the 0.05% level significantly obscured real gradients the SSD surfaces to the point that finding accurate parameters became impossible. Mathematically, this lack of uniqueness is due to the fact that the slope of k_{eff} as a function of the parameters is very small for some solutes and experimental conditions. This is not as much of a problem for fitting the isocratic retention data, such that fitting isocratic data is more robust against the influence of measurement noise when compared to both the basic and *GlobalSearch* approaches to fitting gradient retention data.

Our view is that there are two distinct challenges we face in fitting gradient retention data; 1) lack of “uniqueness’; and 2) inability to simply find the global minimum in a complex fitting landscape. Figure 9 shows several representative fits selected from the results shown in Fig. 11C. Pictured in Fig. 9A-D are the synthetic gradient data (red points) with noise added at $\sigma = 0.5\%$ along with the corresponding fits produced by *GlobalSearch* (black line) for 2,2’-dipyridyl (A), benzonitrile (B), 4-n-butylbenzoic acid (C), and trans-stilbene (D). Additionally, Figs. 9E-H show the corresponding predictions of isocratic k using the parameters obtained from fitting the gradient data (black line) across the range of ϕ=0.0-1.0 compared to the k_{ref} values (red points). In all four examples, the fit of the NK model produced by *GlobalSearch* to the synthetic gradient data resulted in a standard error of the fit that was comparable to or less than the standard error in the gradient retention data introduced by the noise itself, as determined by a comparison via F-test. However, even though the standard errors for each fit were comparable to the standard errors of the data, the MRPE for each fit spanned several orders of magnitude, ranging from $3.92\times10^{-1}\%$ to $1.16 \times 10^{3}\%$. We recognize that such a comparison requires extrapolation of the model to ϕ values outside of the range of conditions experienced by the molecules in the simulated gradient experiments (i.e., $0.05 < \phi < 0.7$; and, some weakly retained analytes will not even experience a large fraction of this range). Reducing the scope of the error calculation to $0.05 < \phi < 0.7$ does reduce the errors substantially (0.2, 2.0, 148, and 36% for 2,2’-dipyridyl, benzonitrile, 4-n-butylbenzoic acid, and trans-stilbene, respectively), but two of them are still much higher than 1%. For some compounds, such as 2,2’-dipyridyl, the fit parameters yielded accurate predictions of isocratic k values across the entire range of ϕ. For other compounds, such as benzonitrile, significant errors in the prediction of isocratic k values were only observed at only one end of the isocratic range. For the others - 4-n-butylbenzoic acid and trans-stilbene – significant prediction errors were observed at both ends of the range of ϕ. These plots demonstrate a distinction between fits where the main problem is that the global minimum has not been found (such as those shown in Fig. 6A/C) and fits where the main problem is a lack of 'uniqueness' in the parameter landscape. *GlobalSearch* is better able to locate the global minimum in the fitting landscape as evidenced by similarity of the standard error of the fit to the standard error of the data. However, even when this algorithm was able to converge upon parameters where the standard error of the fit was comparable to the standard error of the noise in the data, this did not guarantee accurate predictions of isocratic k. Even worse, in practice it is not obvious how one would distinguish between fits that will result in accurate vs. inaccurate isocratic predictions given that the standard errors in both cases are comparable to (or better than) the standard errors of the noisy data itself. For 60 out of the 61 fits in Fig. 8C, the standard errors of the fit were comparable to or better than the standard errors of the data. This suggests that the correlation between gradient data and the underlying NK parameters is fundamentally weaker than that for isocratic data – while it is possible to recover the underlying parameters from gradient data given a sufficiently small level of noise, it is much easier to do so with isocratic data. While this difference is not likely to affect the accuracy of predictions of gradient k_{eff} from gradient data, the impact on the accuracy of predictions of isocratic k can be significant.

5. Conclusions

In this work we have studied factors that affect extraction of retention model parameters from isocratic and gradient and elution retention data. We have used synthetic retention data – modelled after experimental data collected under isocratic reversed-phase conditions for 61 analyte/stationary phase pairs – to enable a detailed investigation of the factors affecting fitting of data to the Neue-Kuss retention model without the complications invariably encountered with experimental data. Following are the principal conclusions drawn from the study.

- Unsupervised fitting of synthetic, noise-free isocratic retention data using a basic trust-reflective region algorithm yields fitting parameters that enable accurate recovery of the original isocratic retention factors. When noise is added to the synthetic data to simulate measurement noise, the accuracy of predictions of isocratic retention factors using the fitting parameters degrades significantly, roughly in proportion to the noise level.

- Unsupervised fitting of synthetic, noise-free gradient elution retention data using the same basic trust-reflective region algorithm yields fitting parameters that cannot consistently accurately predict isocratic retention factors. Adding noise to the synthetic gradient retention data to simulate measurement noise makes the prediction accuracy even worse.
- A good initial guess to initiate fitting using the basic trust-reflective region algorithm improves the predictive accuracy of the resulting retention parameters substantially. However, a significant improvement in performance demands a very high quality guess. For example, starting with an initial guess only 10% different from the known model parameters still produced some errors larger than 1% in isocratic retention factor, even when starting with noise-free gradient retention data, and we are unaware of any current approach that could provide such good initial guesses without considerable experimental effort.
- Using a more sophisticated fitting algorithm that more systematically searches the parameter space for the best solution – *GlobalSearch* in this case – significantly improves the fitting performance for gradient retention data, compared to the use of the basic trust-reflective region algorithm. However, again performance with this approach is not consistent enough to be completely trusted for the purpose of extracting retention model parameters to be used for predicting isocratic retention factors. Our results suggest that this task is challenging for two distinct reasons: 1) the parameter space containing potential model parameters is vast (particularly in the k_w parameter, which spans many orders of magnitude), and in some cases populated by numerous barriers that the fitting algorithm must getting over to find the correct solution – this facet of the problem could be solved using a *GlobalSearch* type of algorithm and a fine parameter grid, at considerable computational expense (e.g., hours per fit on a typical desktop computer); and 2) there is frequently a lack of 'uniqueness' in the parameters obtained from fitting gradient data – that is, there are many combinations of model parameters that lead to fits of similar quality, as measured by the standard error of the fit. This facet of the problem cannot be solved by the search algorithm – it is fundamentally a challenge associated with the nature of the data and the retention model. One possible strategy to alleviate this difficulty is to reparametrize

the model to predict the retention factor at a different mobile phase composition, i.e., instead of k_w (pure water), to a retention factor at a different organic phase composition, as suggested recently by Peris-García et al. [26]. We are currently investigating the potential utility of this approach to address the uniqueness problem.

These results suggest that with current knowledge and retention fitting algorithms it is not possible to consistently obtain retention model parameters that can be used to accurately predict isocratic retention factors from gradient elution retention times. This is the case even with noise-free, synthetic data, where we know the correct answers. Working with experimental data will make the situation worse. If one can tolerate a non-trivial error rate (e.g., more then 5% of results leading to errors in isocratic k >> 1%), then using a thorough search algorithm such as *GlobalSearch* will help improve the likelihood of obtaining useful results. Even so, it would be wise to somehow validate the resulting parameters, perhaps using targeted isocratic experiments.

5. Acknowledgements

T.B. and D.S. acknowledge financial support from the National Science Foundation (CHE-2003734) toward the completion of this study.

References

[1] Kimata, K., Iwaguchi, K., Onishi, S., Jinno, K., Eksteen, R., Hosoya, K., Araki, M., Tanaka, N., Chromatographic characterization of silica C18 packing materials. Correlation between a preparation method and retention behavior of stationary phase. *J. Chromatogr. Sci.* 1989, *27*, 721–728.

[2] Cruz, E., Euerby, M. R., Johnson, C. M., Hackett, C. A., Chromatographic classification of commercially available reverse-phase HPLC columns. *Chromatographia* 1997, *44*, 151–161.

[3] Snyder, L. R., Dolan, J. W., Carr, P. W., A new look at the selectivity of RPC columns. *Anal. Chem.* 2007, *79*, 3254–3262.

[4] Žuvela, P., Skoczylas, M., Jay Liu, J., Bączek, T., Kaliszan, R., Wong, M. W., Buszewski, B., Column characterization and selection systems in reversed-phase high-performance liquid chromatography. *Chem. Rev.* 2019, *119*, 3674–3729.

[5] Grushka, E., Grinberg, N. (Eds), Advances in Chromatography. CRC Press, Boca Raton 2012, pp. 297–376.

[6] Groskreutz, S. R., Weber, S. G., Quantitative evaluation of models for solvent-based, on-column focusing in liquid chromatography. *J. Chromatogr. A* 2015, *1409*, 116–124.

[7] Moussa, A., Lauer, T., Stoll, D., Desmet, G., Broeckhoven, K., Numerical and experimental investigation of analyte breakthrough from sampling loops used for multi-dimensional liquid chromatography. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2020.461283.

[8] Stoll, D. R., Sajulga, R. W., Voigt, B. N., Larson, E. J., Jeong, L. N., Rutan, S. C., Simulation of elution profiles in liquid chromatography − II: Investigation of injection

volume overload under gradient elution conditions applied to second dimension separations in two-dimensional liquid chromatography. *J. Chromatogr. A* 2017, *1523*, 162–172.

[9] Rutan, S. C., Jeong, L. N., Carr, P. W., Stoll, D. R., Weber, S. G., Closed form approximations to predict retention times and peak widths in gradient elution under conditions of sample volume overload and sample solvent mismatch. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2021.462376.

[10] Gritti, F., Gilar, M., Hill, J., Mismatch between sample diluent and eluent: Maintaining integrity of gradient peaks using *in silico* approaches. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2019.460414.

[11] López-Ureña, S., Torres-Lapasió, J. R., García-Alvarez-Coque, M. C., Enhancement in the computation of gradient retention times in liquid chromatography using root-finding methods. *J. Chromatogr. A* 2019, *1600*, 137–147.

[12] Tyteca, E., Périat, A., Rudaz, S., Desmet, G., Guillarme, D., Retention modeling and method development in hydrophilic interaction chromatography. *J. Chromatogr. A* 2014, *1337*, 116–127.

[13] Dolan, J. W., Lommen, D., Snyder, L. R., DryLab computer simulation for high-performance liquid chromatographic method development. II. Gradient elution. *J. Chromatogr. A* 1989, *485*, 91–112.

[14] Pirok, B. W. J., Pous-Torres, S., Ortiz-Bolsico, C., Vivó-Truyols, G., Schoenmakers, P. J., Program for the interpretive optimization of two-dimensional resolution. *J. Chromatogr. A* 2016, *1450*, 29–37.

[15] den Uijl, M. J., Schoenmakers, P. J., Schulte, G. K., Stoll, D. R., van Bommel, M. R., Pirok, B. W. J., Measuring and using scanning-gradient data for use in method optimization for liquid chromatography. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2020.461780.

[16] Kensert, A., Collaerts, G., Efthymiadis, K., Desmet, G., Cabooter, D., Deep Q-learning for the selection of optimal isocratic scouting runs in liquid chromatography. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2021.461900.

[17] Snyder, L. R., Dolan, J. W., High-Performance Gradient Elution: The Practical Application of the Linear-Solvent-Strength Model. John Wiley, Hoboken, NJ 2007.

[18] Neue, U. D., Kuss, H.-J., Improved reversed-phase gradient retention modeling. *J. Chromatogr., A* 2010, *1217*, 3794–3803.

[19] Kromidas, S. (Ed.), The HPLC Expert II: Find and Optimize the Benefits of Your HPLC/UHPLC. Wiley-VCH, Weinheim 2017, pp. 101–170.

[20] Bos, T. S., Niezen, L. E., den Uijl, M. J., Molenaar, S. R. A., Lege, S., Schoenmakers, P. J., Somsen, G. W., Pirok, B. W. J., Reducing the influence of geometry-induced gradient deformation in liquid chromatographic retention modelling. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2020.461714.

[21] Tyteca, E., Guillarme, D., Desmet, G., Use of individual retention modeling for gradient optimization in hydrophilic interaction chromatography: Separation of nucleobases and nucleosides. *J. Chromatogr. A* 2014, *1368*, 125–131.

[22] Navarro-Huerta, J. A., Gisbert-Alonso, A., Torres-Lapasió, J. R., García-Alvarez-Coque, M. C., Testing experimental designs in liquid chromatography (I): Development and validation of a method for the comprehensive inspection of experimental designs. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2020.461180.

[23] Vaast, A., Tyteca, E., Desmet, G., Schoenmakers, P. J., Eeltink, S., Gradient-elution parameters in capillary liquid chromatography for high-speed separations of peptides and intact proteins. *J. Chromatogr. A* 2014, *1355*, 149–157.

[24] Cabooter, D., Song, H., Makey, D., Sadriaj, D., Dittmann, M., Stoll, D., Desmet, G., Measurement and modelling of the intra-particle diffusion and b-term in reversed-phase liquid chromatography. *J. Chromatogr. A* DOI: 10.1016/j.chroma.2020.461852.

[25] Vaast, A., Tyteca, E., Desmet, G., Schoenmakers, P. J., Eeltink, S., Corrigendum to "Gradient-elution parameters in capillary liquid chromatography for high-speed separations of peptides and intact proteins" [J. Chromatogr. A 1355 (2014) 149–157]. *J. Chromatogr. A* 2014, *1366*, 137.

[26] Peris-García, E., Ruiz-Angel, M. J., Baeza-Baeza, J. J., García-Alvarez-Coque, M. C., Comparison of the fitting performance of retention models and elution strength behaviour in hydrophilic-interaction and reversed-phase liquid chromatography. *Separations* 2021, *8*, 54.

Figure Captions

Figure 1. Description of fitting process and error evaluation. The input data for box 1 was collected using 61 analyte/stationary phase pairs. All parameter sets output in box 2 had R^2 values greater than 0.999 for the isocratic NK model. For the steps where *lsqnonlin* was used, the function tolerance was set to 1×10^{-10}, the number of function evaluations was set to 100,000, and the number of function evaluations was set to 100,000. When *GlobalSearch* was used, the number of trial data points was set to 100,000 and the limits on each parameter were: $k_w =$ 1.0-1.0× 10^9, $S_1 =$ 5.0-400, and $S_2 =$ 0.05-15. The numbers in parentheses are referred to as box 1, etc. in the text.

Figure 2. Performance of unsupervised fitting algorithm on gradient data (Table 1, Row #5). Synthetic gradient retention data for fitting were generated as described in Fig. 1 (box 8) and fit as described in Fig. 1 (box 11a), while data shown for comparison was calculated as described in Fig. 1 (box 12). The initial guess in Fig. 1 (box 10a) was chosen to be $k_w = 1.0$, $S_1 = 1.0$, and $S_2 = 1.0$. Example plots of ln (k) vs. ϕ are shown for 2,2'-dipyridyl (A) and benzonitrile (B), where the isocratic k_{ref} values are displayed as the red points, and the k_{fit} values calculated from the fit of the synthetic gradient data are shown as the black line. The percent difference between k_{ref} and k_{fit} (D) are shown for all 61 compounds, as well as a plot of k_{fit} vs. k_{ref} (C).

Figure 3. Distribution of average of the absolute value of the percent errors between isocratic k_{ref} and isocratic k_{fit} after fitting noise-free gradient data with an unsupervised algorithm using several initial guesses. The histograms shown contain errors for all 61 sets of parameters. Fitting data were generated as described in boxes 8 and 11a of Fig. 1, with percent errors calculated as described in box 12. The initial guess in box 10a was chosen using the equation $x_0 = \alpha \times x_{ref}$, where x_{ref} is the parameter set obtained in box 2 and α is a multiplier. α values for each plot were: $1\text{x}10^{-8}$ (a), 1 (b), 1.1 (c), and 2 (d).

Figure 4. Plot of sum of squared differences (SSD) between gradient k_{eff} values calculated using either 'reference' NK parameters (Fig. 1, box 2)), or a set of parameters indicated by a point in the three-dimensional space, for ten different gradient times. Reference parameters were: A) $k_w = 2.077 \times 10^8$, $S_1 = 199.5$, and $S_2 = 7.297$ for amitriptyline (a) on SB-C18; and B) $k_w = 63.89$, $S_1 = 63.98$, and $S_2 = 7.344$ for 2,2'-dipyridyl. Gradient parameters were

$\phi_0 = 0.05$, $\Delta\phi = 0.65$, $t_d = 0$ min, and $t_0 = 0.1$ min. Effective retention factors were calculated using Eq. 3.

Figure 5. Plot of average of the absolute values of the percent difference between isocratic k values calculated using either 'reference' NK parameters (Fig. 1, box 2)) (k_{ref}), or a set of parameters obtained by fitting gradient k_{eff} values using a point in the three-dimensional space as an initial guess, for ten different ϕ values between ϕ_{lower} and ϕ_{upper} (k_{fit}). Reference parameters were $k_w = 2.077 \times 10^8$, $S_1 = 199.5$, and $S_2 = 7.297$ for amitriptyline (A) on SB-C18 and $k_w = 63.89$, $S_1 = 63.98$, and $S_2 = 7.344$ for 2,2'-dipyridyl (B). Gradient conditions were the same as those used in Fig. 5. Noise-free synthetic gradient retention data were fit with *lsqnonlin* in MATLAB using the trust-region-reflective algorithm with the number of iterations set to 100,000, the number of function evaluations set to 100,000, and the function tolerance set to 1×10^{-10}.

Figure 6. Fits obtained after applying the basic unsupervised fitting algorithm (Fig. 1 (box 11a)) to synthetic gradient retention data (Fig. 1 (box 8)) for 2,2'-dipyridyl using several different initial guesses as the starting point. Plots A-C show the final fit (solid line) to the gradient retention data (red points) for each starting point, while plots D-F show comparisons of the corresponding isocratic predictions (k_{fit}; black line) to k_{ref} (red points) (Fig. 1 (box 12)). The reference parameters (and those converged to in plot (B)) were $k_w = 63.89$, $S_1 = 63.98$, and $S_2 = 7.344$; the parameters converged to in plot (A) were $k_w = 1.116 \times 10^9$, $S_1 = 2.318\times 10^3$, and $S_2 = 87.38$; the parameters converged to in plot (C) were $k_w = 1.132 \times 10^9$, $S_1 = 2.321\times 10^3$, and $S_2 = 87.43$.

Figure 7. Distributions of average of the absolute value of the percent differences (Fig. 1 (box 7)) between k_{ref} and k_{fit} after fitting noisy gradient retention data (Fig. 1 (box 8/9)) using an unsupervised algorithm (Fig. 1 (box 10a/11a)) for all 61 sets of NK parameters and different noise levels: A) $\sigma = 0\%$; B) $\sigma = 0.05\%$; C) $\sigma = 0.5\%$. The initial guess in Fig. 1 (box 10a) was chosen to be $k_w = 1.0$, $S_1 = 1.0$, and $S_2 = 1.0$.

Figure 8. Distributions of average of the absolute value of the percent differences (Fig. 1 (box 7)) between k_{ref} and k_{fit} after fitting noisy gradient retention data (Fig. 1 (box 8/9)) using a the

GlobalSearch algorithm (Fig. 1 (box 10b/11b)) for all 61 sets of NK parameters and different noise levels: A) σ = 0%; B) σ = 0.05%; C) σ = 0.5%.

Figure 9. Fits obtained after applying the *GlobalSearch* fitting algorithm (Section 3.2.2.2, and Fig. 1 (box 11b)) to synthetic gradient retention data with 0.5% relative noise added (Fig. 1 (box 8)) for several compounds. Plots A-D show the fit produced by *GlobalSearch* (solid line) to the synthetic gradient data (red points), while plots E-H show comparisons of the corresponding isocratic predictions (k_{fit}; black line) to k_{ref} (red points) (Fig. 1 (box 12)). Panels A and E correspond to 2-2'dipyridyl; B and F correspond to benzonitrile; C and G correspond to 4-n-butylbenzoic acid; and D and H correspond to trans-stilbene. For each fit shown in A-D, the standard error of the fit was either equivalent to the standard error of the noise (determined by F-test) or was lower. The mean residual percent errors for plots E-H were 0.392, 3.49, 1.16×10^3, and 39.1%, respectively when calculated over the range $0 < \phi < 1.0$. Reducing the range to $0.05 < \phi < 0.7$ reduces the errors to 0.2, 2.0, 148, and 36%, respectively.